# Predicting withdrawal based on student information and VLE engagement in the Open University

## University of Suffolk

## Jack Milton

## Table of Contents

[Table of Contents 1](#_Toc161941619)

[Abstract 2](#_Toc161941620)

[Introduction 2](#_Toc161941621)

[Methods and Experimental Setup 3](#_Toc161941622)

[Cleaning and Preparing Data 3](#_Toc161941623)

[Feature Engineering 4](#_Toc161941624)

[Merge 4](#_Toc161941625)

[Data Pre-processing 4](#_Toc161941626)

[Statistical Methods 5](#_Toc161941627)

[Machine Learning Methods 7](#_Toc161941628)

[Results and Discussion 9](#_Toc161941629)

[Conclusion 12](#_Toc161941630)

## Abstract

This project utilises the Open University Learning Analytics dataset (OULAD) to explore the predictive potential of student withdrawal from courses based on their personal information and virtual learning environment engagement (OULAS, 2017). By using machine learning models, the study aims to shed light on the underlying reasons for student withdrawals at the Open University. Despite encountering challenges, such as imbalanced data leading to lower F1-scores, the analysis provides valuable insights into potential correlations between student characteristics and withdrawal behaviours. These findings hold promise for implementing strategies to prevent future student withdrawal and enhance overall student success at the Open University.

## Introduction

A previous research study attempted to predict withdrawal from higher education using demographic, psychological and educational measures. The study identified demographic features as significant predictors for the retention of students. It was also found that academic features, such as previous education results and low university entry requirements, influenced students’ decisions on whether to withdraw. However, this study has highlighted limitations. Although the logistic regression model was shown to be significantly predictive, the classification figures show that the model could not correctly classify between classes very well. Furthermore, while certain features may explain a portion of the variance in whether a student will withdraw, the predictive power may be limited. Therefore, there is a need for a different approach that considers other factors that could contribute to student withdrawal, such as students' interaction with their studies encapsulating early engagement (Charlton et al, 2006).

In this study, I am attempting to solve a crucial question: Could we predict whether a student will withdraw from a course based on their personal information and how they interact with the virtual learning environment? This inquiry aligns with the mission statement(s) of the Open University, which is dedicated to fostering student success. The use of data analysis, as provided by this study, will allow greater accuracy in predicting student withdrawal. This will then greatly impact the institution and its student body, potentially leading to improved retention rates and the optimised allocation of resources through targeted interventions. This investigation lies in the recognition that early identification of students at risk of withdrawal is essential for implementing proactive intervention strategies within higher education. By pinpointing individuals who may be considering dropping out at an early stage, universities can deploy timely support mechanisms tailored to each student's specific needs. This approach enhances academic outcomes and nurtures a supportive, inclusive learning environment to sustain student success. Furthermore, the identification of students in need aligns with the Open University's core values of accessibility and inclusivity, ensuring that all learners, regardless of their background or circumstances, have equitable opportunities to thrive academically (REF).

To guide the exploration into this complex issue, the following research questions were devised:

*Can we develop accurate predictive models for student dropout using personal information and engagement metrics within the virtual learning environment?*

*What primary determinants contribute to student withdrawal, as evidenced by the dataset?*

These questions serve as the foundation for my investigation, aiming to uncover insights that could change how universities approach student retention and support mechanisms.

## Methods and Experimental Setup

### Cleaning and Preparing Data

The first part of my process, in fulfilling this brief, was to check for any missing values across the data frames. I noticed a substantial number of missing values ‘week\_from’ and ‘week\_to’; over 83% of the data was missing, so I decided to drop these from the data frame. This is an important part of the process, to include at the early stages, to ensure clean data, meaning that the data from which results will be drawn will be of high quality. I removed rows where date registration was missing as it inferred that the student had not registered and, therefore, could not withdraw. The missing values in the IMD band were imputed with the mode for each region, which iterated over each region to increase accuracy.

### Feature Engineering

The datasets VLE and student VLE were merged using common keys to reflect the correlation between these two groups. This is an example of feature engineering wherein the existing vectors are adapted and/or replaced (engineered) to ensure accuracy in the results (Heaton, 2016). The feature engineering occurred over two features, one encapsulating early engagement via the total sum of clicks made by a student per 14 days. Secondly, I replicated this with activity: the most clicked activity over 14 days.

I had to fill the missing values with 0 and NaN since some students did not interact with the VLE. This would negatively impact my algorithm; therefore, they had to be adapted and/or removed.

### Merge

Following this, I merged the data frames together: the VLE merged, student info, student registration and courses to create one data frame. Once merged, I dropped ID columns since they had data leakage to the target variable. As I will not have this information for predicting in the future, this needed to be removed. I dropped activity type, sum click, and date because I had used these features within the feature engineering to avoid multicollinearity.

### Data Pre-processing

I converted number of previous attempts and module presentation length into categorical variables as it made more sense than their previous numerical variable type due to their having a limited consistent number of values. I analysed the unique counts in most frequent activity, resulting in at least four groups needing to be grouped together in aid of being recognised by machine learning models. Then, I converted the column final result column into a binary column named withdrawn, therefore any value that wasn’t withdrawn would be converted into 0 for not withdrawn. This was vital as I was using classification algorithms which predict categorical variables as opposed to a regression algorithm that expects continuous data.

I sorted the columns into ordinal, nominal and numerical columns. Through analysing the data, I was able to discern which grouping would be best. For example, I wanted to preserve the integrity of some of the categorical columns as they already had a notion of order thus making them ordinal **(Ross Nelson, 2023 p. 232)** The IMD band feature represents deciles of the Index of Multiple Deprivation, with '0-10%' being the most deprived and '90-100%' being the least deprived. In many contexts, a higher value represents a higher level of something. So, to align with this common interpretation, the mapping is done in reverse order: '0-10%' (most deprived) is mapped to 9, and '90-100%' (least deprived) is mapped to 0. Thus, a higher value in the IMD band feature represents a higher level of deprivation, which might make the model's behaviour more intuitive to understand. I then sorted them into their corresponding transformers: standard scaler, a pre-processor which standardises the features by removing the mean and standardising to the unit variants **(p. 317)**, for numerical columns. I used one hot encoder and ordinal encoder for the categorical variables. The use of transformers is vital to ensure the machine learning model can correctly interpret the data.

### Statistical Methods

With the statistical testing the motivation was to discern whether there was a statistical significance between the variables and the target variable. Therefore, I devised some hypotheses:

*= There is no significant difference between categorical variables and withdrawn.*

*= There is a significant difference between categorical variables and withdrawn.*

To do this I checked the assumptions to see if it met the chi-square test (**Currell and Dowman, 2009, p. 332** ) using the chi-square contingency table and it did meet the assumptions. Therefore, I ran the test and found all categorical variables had a statistically significant relationship with the target variable thus rejecting null hypothesis.

Next, was the numerical variables. I needed to test the assumptions first, so I devised another hypothesis for a Shapiro-Wilk test to test for normality:

*= The data has a normal distribution*

*= The data does not have a normal distribution*

I found non-normalisation amongst all variables, so I rejected the null hypothesis that they were normally distributed. The assumption of normality was not met which allowed me to proceed to the Mann-Whitney U test **(p.299)**. From this, I continued to my third hypothesis:

*= There is no significant difference between numerical variables and withdrawn.*

*= There is a significant difference between numerical variables and withdrawn.*

I found that there was a statistically significant difference with all numerical variables and the target variable, therefore, rejected null hypothesis. Although, these show significant difference this is not causation because the statistical test can only show correlation as it does not account for any other variables which may influence cause.

Exploring my variables was first done using box plots to perform outlier analysis for the numerical columns. To address the outliers, I capped them to reduce the impact of the outliers on my model. This is important because it can lead to bad performance through skewing of the data. I then performed a univariate analysis which examined the variables individually which enabled a better comprehension of the data. For the numerical columns I found that early engagement was skewed to the left and date registration had a standard distribution **(Currell and Dowman, 2005, p. 299)**

I then performed bivariate analysis **(Freedman, 2009, p. 215)** which is analysing the variables with the target variable withdrawn. This also improved my comprehension since it allowed me to identify that over half of those who did not engage with the VLE withdrew in comparison to those who did.

I also created a correlation matrix which uses the Pearson Correlation Coefficient which measures the linear relationship between the variables **(Currell and Dowman, 2005, p. 316).** This is a useful tool in getting a different representation of the variables which lead to a higher comprehension of the data from which results and analysis can occur.

### Machine Learning Methods

Using the previously created encoders, they were input into a ColumnTransformer, which is the pre-processor for my models. I then created a pipeline for each of the models, which included the pre-processor of my variables and the classifier of the type of model being used, e.g. Logistic regression, RandomForest; within the classifier would be the attached hyper-parameters, which I would subsequently tune using grid-search. Also, inside the pipeline would be the technique of handling the imbalance, e.g. SMOTE, RandomUnderSampler. I then defined the target variable and features.

= withdrawn

= nominal, ordinal and numerical features

This means I am separating the target variable in the data frame from the rest of the features. Then, I split the data into training and test sets. If I were using a technique, this would be applied in the next line of code; if not, I would go straight to fitting the model. The probability of the positive class for the training set would then be predicted; then I classified the instances based on the decision threshold, then predict the probabilities of the positive class, and classify the instances based on the decision threshold, leading to model evaluation.

First, I used the multi-layer perception classifier, which is a type of a neural network model. This is an artificial neural network that consists of multiple layers of neurons. Each neuron in a layer is connected to every neuron in the subsequent layer, you could also consider the neuron as a node. I was able to use the hyper-parameters to change this. One of which was the hidden layer sizes, which is the number of neurons used. I used 50 as a result from using a grid search.

Second, I used logistic regression:

represents the probability of the target variable being 1 given the features .

is the feature matrix.

is the weight vector (coefficients).

is the bias term (intercept).

is the base of the natural logarithm.

(Bishop, and Nasrabadi, 2006.  p. 738)

with different techniques to enable me to observe and analyse the optimum solution for handling imbalanced data. I had to handle the imbalanced data accordingly since the not withdrawn class (0) of the dataset is double that of the withdrawn class (1). This is problematic for my purposes of predicting the likelihood of withdrawal since it has lower instances, which will affect accuracy. Therefore, I had to apply different techniques in the pipeline of my models; using classifiers without these measures would not address the issue of the imbalanced dataset and cause poor performance in the minority class (1) (Kotsiantis et al. 2006, p. 25).

I used different pipelines for different techniques to ensure suitability. I devised a line of code which would enable me to achieve a balance between recall and precision, which was a decision threshold. This threshold is automatically set at 0.5, but I found for most of the models, it worked optimally between 0.6-0.605. I tuned these models using grid search with various hyper-parameters. Also, I used Gradient Boosting Classifier, Support Vector Machines and Random Forest to which the performance was not as successful as logistic regression. This could be due to model complexity, considering logistic regression is a simpler model compared to the others. Furthermore, due to my machine's limitations, the models that didn’t perform very well were highly dependent on their hyperparameters and the tuning, which was computationally too expensive for my machine and, therefore, was only tuned to a very basic level.

The classification report was used to print the results of the training set and the test set. I was most concerned about the F1 score, which is the average between precision and recall. I used the F1 score as the metric due to the imbalanced dataset, as it is the harmonic mean of precision and recall and provides a balanced measure of the model’s performance on both classes **(Grus, 2019, p. 159).** Improvement was the goal; hence when tuning grid searches, I used the scoring metric as F1 macro. Cross-validation was another way in which I evaluated the models. I used stratified k-fold cross-validation because it is essential when the class distribution is uneven. This helps prevent bias in the model evaluation.

## Results and Discussion

Training Set Performance:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Class 0 | 0.78 | 0.84 | 0.81 | 4497 |
| Class 1 | 0.58 | 0.49 | 0.53 | 2013 |
| Macro avg | 0.68 | 0.67 | 0.67 | 6510 |
| Weighted avg | 0.72 | 0.73 | 0.73 | 6510 |

Test Set Performance:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Class 0 | 0.79 | 0.84 | 0.81 | 4497 |
| Class 1 | 0.58 | 0.49 | 0.53 | 2013 |
| Macro avg | 0.68 | 0.67 | 0.67 | 6510 |
| Weighted avg | 0.72 | 0.73 | 0.73 | 6510 |

These are the results from the best-performing model, Logistic Regression, using SMOTE to handle class imbalance. By setting the threshold manually to balance precision and recall, in this instance, I used 0.605; the best achievable f1-score was 0.6717220229120069, which means my model is correctly predicting likelihood of student withdrawal 67% of the time. The model performed well on Class 0 due to more instances of that class, which allowed the model to find more true positives. From the AUC-ROC curve I plotted, I received a result of 0.7366871837393169, which suggests the model can distinguish between classes reasonably well.

Results from cross-validation:

Cross-validation F1 scores: [0.64297706 0.46766399 0.47885454 0.47977257 0.61337053]

Mean cross-validation F1 score: 0.5365277358299325

Standard deviation of cross-validation F1 score: 0.07553255546407642

From this, we can tell that the model’s performance may be somewhat unstable across different subsets of the data due to the lower mean f1-score. Also, the standard deviation score indicates that there is some variability across different folds. The discrepancy between these results and the original scores suggests there is a degree of overfitting; this could be due to the model being too complex and fitting too closely to the specific characteristics of the training data, therefore reducing its ability to generalise to unseen data (REF).

The inability to use more powerful models, that can handle more dimensions, was a limitation. due to the computational expense of tuning and running the models and the restrictions of logistic regression being more simplistic but less computationally expensive to tune. The computational expense was the main contributor to this issue, meaning that I relied on a more simplistic but less computationally expensive model. Handling imbalance was also a prominent issue, drastically reducing performance. Therefore, I turned my attention to finding the best possible balance so the model could distinguish between the classes.

Even with the limitations, I was able to derive meaningful insights from this model using the feature coefficients from the model.

Coefficients represent the change in log odds of the outcome for a unit in the feature; the magnitude, which is the absolute value, indicates the strength of the relationship between the feature and the outcome. Larger values mean that the feature has a stronger impact on the outcome; values of the coefficients can be impacted by the scale of individual features, whilst accounting for the previously scaled and normalised our features in the pre-processor. We can interpret these features by looking at the positive and the negative and what they correlate to.

Positive coefficients correspond to the positive class (Class 1/Withdrawn), therefore making them more likely to withdraw. Negative coefficients react the opposite way as the log odds of the positive class decrease, which then increases the negative class (Class 0/Not Withdrawn), thus making them more likely to not withdraw.

Here are the top fifteen features that are sorted by the absolute value to determine importance.

|  |  |
| --- | --- |
| Features | Coefficients |
| code\_module\_GGG | -1.076493 |
| code\_module\_CCC | 0.975591 |
| most\_freq\_activity\_NULL | 0.961044 |
| code\_module\_DDD | 0.481408 |
| code\_module\_EEE | -0.405342 |
| code\_module\_AAA | -0.377865 |
| region\_Ireland | -0.355448 |
| num\_of\_prev\_attempts | -0.354505 |
| most\_freq\_activity\_htmlactivity | 0.343594 |
| early\_engagement | -0.323691 |
| most\_freq\_activity\_quiz | -0.313920 |
| most\_freq\_activity\_oucontent | -0.302314 |
| studied\_credits | 0.298436 |
| region\_Wales | -0.292433 |
| most\_freq\_activity\_forumng | -0.287306 |

From this, I can infer that students studying different courses are more likely and less likely to withdraw. However, due to the Data Protection Act, I am limited in the data I can input and therefore cannot offer specified predictions. If I had access to the names of the courses, I could potentially infer that students studying STEM subjects are more likely to withdraw, which could be due to higher cognitive strain of the courses and/or more complicated workloads.

The feature early engagement encapsulates students who engage more with the virtual learning environment, and the results demonstrate that the more the student engages: the less likely they are to withdraw. Feature ‘most\_freq\_activity\_NULL’ shows that students who do not interact with the VLE at all have a higher likelihood of withdrawal due to their strong relationship with the positive class. Students who are studying more credits, therefore suggesting having a higher workload, are more likely to withdraw. Students who are studying in Ireland are also less likely to withdraw. This could be due to socioeconomic factors, since studies show that Ireland is significantly wealthier than Britain; a better economy could suggest a better standard of living and higher likelihood of student financial stability (Gosling, 2023). Thus, suggesting if a student is more financially stable, they may have more focus on their studies. However, this is conjecture based on the data and study cited above.

## Conclusion

Despite encountering challenges such as low to average performance with the models, the study provides valuable insights into the correlations between student characteristics and withdrawal behaviours. The results of the study indicate that, to a certain extent, predicting student withdrawal is possible with the model correctly predicting student withdrawal only 67% of the time. The study identified several factors that contribute to student withdrawal, such as: certain course modules, previous education, differing regions, and VLE interaction.

Overall, while the study provides valuable insights into student withdrawal, it also reflects the complexity of this task, its issues, and limitations. Therefore, future studies could explore more sophisticated methods by employing more computationally expensive models that could perhaps handle the class imbalance more efficiently to improve overall performance.

**Bibliography**

(Bishop, C.M. and Nasrabadi, N.M., (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.)

Charlton, J. P., Barrow, C. and Hornby-Atkinson, P. (2006) "Attempting to predict withdrawal from higher education using demographic, psychological and educational measures.", *Psychology: Journal Articles (Peer-Reviewed)*. 2.[Online]. Available: http://digitalcommons.bolton.ac.uk/psych\_journalspr/2

Currell, G. and Dowman, A. (2009) *Essential mathematics and statistics for science*. 2nd edn. West Sussex: Wiley-Blackwell.

Freedman, D. A. (2009) (2nd edn). *Statistical models : theory and practice*. Leiden: Cambridge University Press.

Gosling, P. (2023). ‘Who is better off: the Irish, the Northern Irish or the British? A regional economic comparison’, *Irish Studies in International Affairs*, vol. 34 (2) p. 186-221. DOI: <https://doi.org/10.1353/isia.2023.a908962>.

Heaton, J. (2016). "An empirical analysis of feature engineering for predictive modelling," *SouthEast con*. Norfolk, VA, USA. 09.07.2016. doi: 10.1109/SECON.2016.7506650. Kotsaintis, S., Kanellopoulos, D., and Pintelas, P. (2006). ‘Handling imbalanced datasets: a review’,in *GESTS international transactions on computer science and engineering*. 30(1) [Online]. Available: <https://d1wqtxts1xzle7.cloudfront.net/40045253/0c960517fefa59fa6b000000.pdf20151115-68247-1ohdc8j-libre.pdf?1447645886=&response-content-disposition=inline%3B+filename%3DHandling_imbalanced_datasets_A_review.pdf&Expires=1710769714&Signature=Pk3tsv3VrsQsxHon0Z0qso79xcGk9JwwFU-i~K~GI-gXXbyswDSqW7wwCG2KTMaUpj0NNgk5OaVGjIA8MYI9UwUIU8LOdG7xmO7jnVOk~ZwORPbTDI0iIW2lzYW3umq04vfFuhTY~ZaI65BDYJ5YNDTfB6vcWk8pV6i2aFHuHav94a14hIkJ3kZaiOAAV9HAk8MxCMsimNL-IE7hOtvyi-FQA-XPFZhS9crPDfzSladCE2Uxgoqo-jceJloE3kxTALRliY3Iyylz4Y2Vv2~xVXFxrg6y6sJKoQxhsQLOdVmoRPHHKugFKhs2lA9OaIo7GecYFpJrnZTvcl9WCQEdxw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA> pp. 25-36.

Ross Nelson, A. (2023). *Confident data science.* London: Kogan Page Ltd.