970G1: Data Science Research Methods

Report (4000 Words) A1 Week 1

Word Count: 3961

Contents

Introduction

Background

The Open University (OU) offers accessible higher education in the UK, with modules starting in February and October. Each module includes Tutor Marked Assessments (TMA), Computer Marked Assessments (CMA), and Final Exams, which contribute to final grades. OU recently acquired a Virtual Learning Environment (VLE) to improve the student learning experience. VLEs collate module resources and engagement methods into a digital platform, allowing students to access materials and complete assessments remotely. Widely used in the UK, popular platforms include Moodle, Blackboard, and Canvas (Mosley, 2024).

The Open University Learning Analytics Dataset (OULAD)

OULAD includes 7 datasets: `assessments, courses, studentassessment, studentinfo, studentregistration, studentvle & vle`. These cover 7 modules across 4 presentation periods (2013–2014), involving 28,785 students.

3 datasets (`courses, assessments & vle`) detail course materials (e.g., *code_module, assessment_type*), while the other 4 (`studentassessment, studentinfo, studentregistration, studentvle`) provide anonymised student demographics (e.g., *gender, age_band*), VLE engagement (*sum_click*), and registration details (*date_registration, date_unregistration*).

Together, these datasets enable analysis of factors influencing student grades (e.g., demographics, VLE engagement) and the VLE's impact on academic outcomes, discussed further in the next section.

Research Hypotheses

The two main research questions being investigated by this report are:

1. Is the VLE improving students' grades?
2. Can we predict students' grades?

Is the VLE Improving Students' Grades?

VLE engagement could be defined in a couple of ways: the average number of clicks per student and, using the *date* (from `studentvle`) a student interacted with a particular VLE material, we can determine the number of unique days a student interacts with VLE materials. These measures can be combined into a single VLE engagement measure.

Instead of using the final grade (*Pass, Fail, Distinction*) as a measure of students' scores, we can use the average assessment score (0-100) as the dependent variable. Accordingly, our hypotheses for this research question are as follows:

$\mathbf{H_a}$: **If** the VLE is improving students' grades, **then** students with "high" VLE engagement should have significantly higher average assessment scores than students with "low" VLE engagement.

$\mathbf{H_0}$: **If** the VLE is not improving students' grades, **then** the average assessment scores for students with "high" VLE engagement will not be significantly greater than the average assessment scores of students with "low" VLE engagement.

Can we Predict Students' Grades?

Using features related to student demographics, VLE engagement, and course performance, a statistical model to predict final grades can be constructed. Instead of average assessment scores, *final_result* is used for practical relevance and to avoid complications with assessment weightings, discussed during exploratory data analysis. The model classifies whether a student passes or fails their module and helps explore how these features influence the likelihood of passing.

Data Cleaning & Wrangling

After inspecting the data, notable points emerged, each addressed during data cleaning. Interpretation was supported by information from Kuzilek et al. (2017).

Null or Inappropriate Values

Whilst no null values were found across the 7 datasets, 7 columns contained "?" indicating unknown values. Rows with "?" in the *score* column of `studentassessment` were excluded, as score is important for the analyses. Other "?" occurrences were excluded only for specific visualisations (e.g., *imd_band* histogram), but not for the main analyses. 173 rows were removed from `studentassessment`.

Duplicates

Using the following unique identifiers for each data set, `studentinfo`, `studentregistration` and *studentvle* contained duplicate rows:

Table 1: Unique Identifiers Used for Each Data Set

Data Set	Unique Identifiers
assessments	id_assessment
course	code_module, code_presentation
studentassessment	id_student, id_assessment
studentinfo	id_student
studentregistration	id_student
studentvle	id_student, id_site, code_presentation, code_module, date
vle	id_site

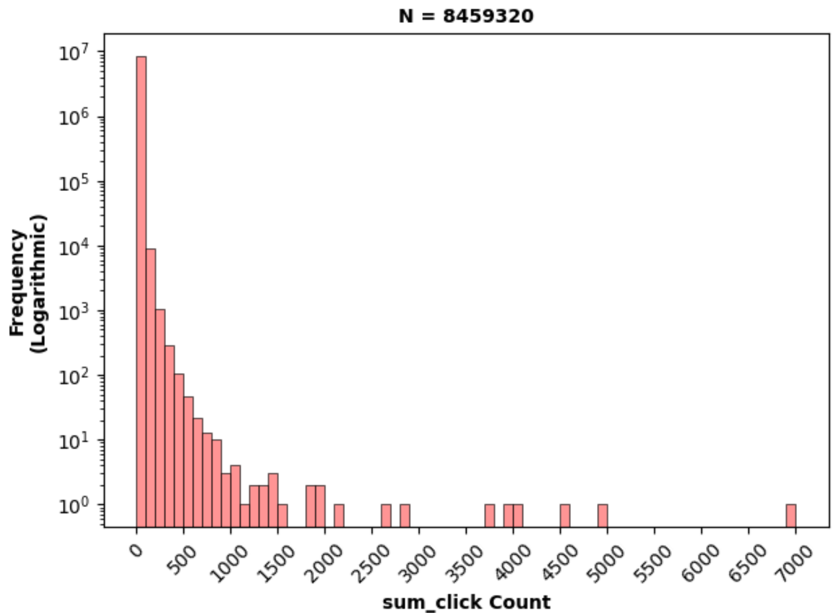
Upon inspection, it was determined that duplicates in `studentinfo` and `studentregistration` held important data, such as module attempts and registration changes. Excluding these rows would lose this information, so they were retained.

For `studentvle`, duplicate rows showed multiple entries for a student's engagement with the same VLE material on a given date. Instead of excluding these, I aggregated their *sum_click* counts, reducing rows from 10,655,280 to 8,459,320 while ensuring each *sum_click* count reflected a student's engagement with one material on one date.

Unexpected Ranges

3 numeric columns had unexpected ranges: *date_submitted* from `studentassessment` had a minimum of -11, *studied_credits* from `studentinfo` had a maximum of 655, and *sum_click* from `studentvle` had a maximum of 6,977. There was no justified reason to exclude rows for the first two. **Figure 1** shows the distribution of *sum_click* on a logarithmic scale.

Figure 1: Histogram of *sum_click* Counts, Using a Logarithmic Scale



Few *sum_click* counts exceed 1,000. **Appendix A** shows no specific students or VLE materials overrepresented these "high" counts. It's unlikely these counts represent genuine user input, as we would expect more students to record similar counts if the VLE required over 1,000 clicks. Therefore, I excluded *sum_click* counts above 1,000, removing 26 rows.

Weighted Scores

Module scores are the sum of each weighted score from module assessments. To calculate this, I created a new column detailing the weighted score for each student by performing a left join on `studentassessment` with the *weight* and *id_assessment* columns from assessments. I then added a *weighted_score* column in the `studentassessment` dataset by multiplying *score* and *weight*.

Exploratory Data Analysis

Student Demographics

Figure 2: Histogram and Bar Charts of Age, Gender and Qualification Distributions

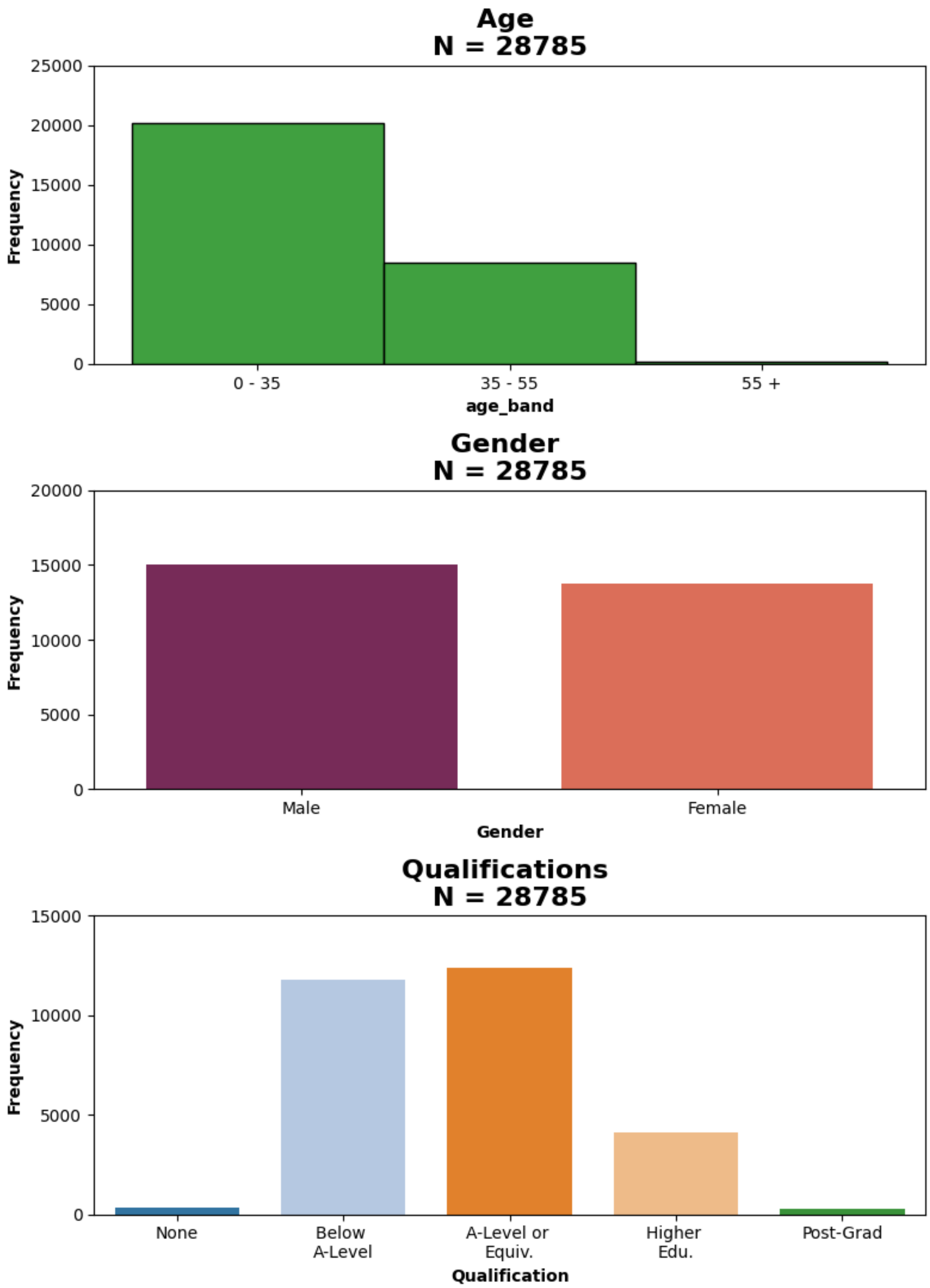
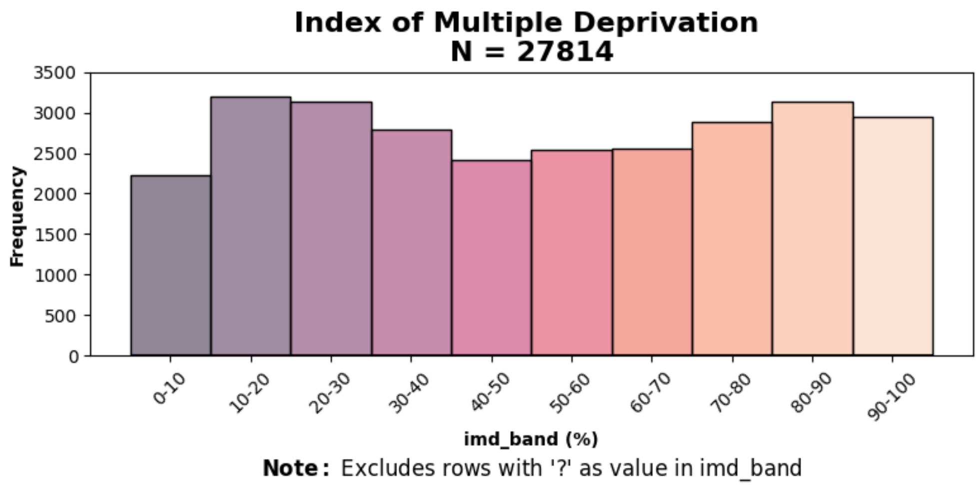
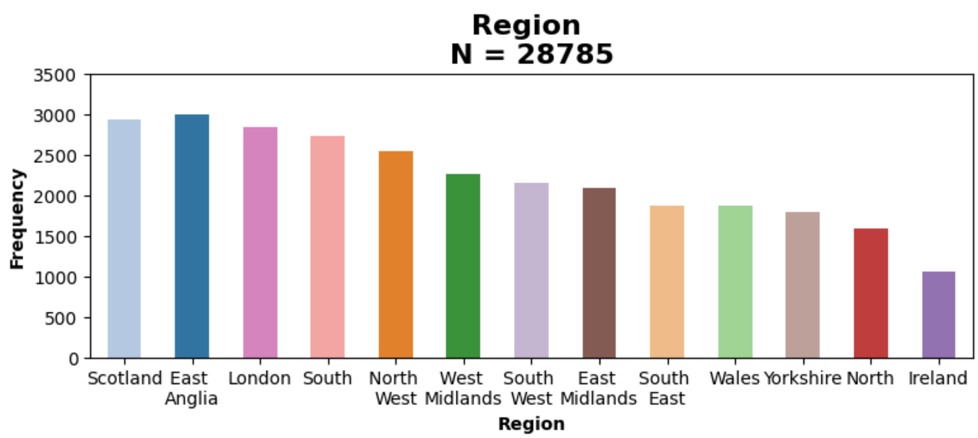
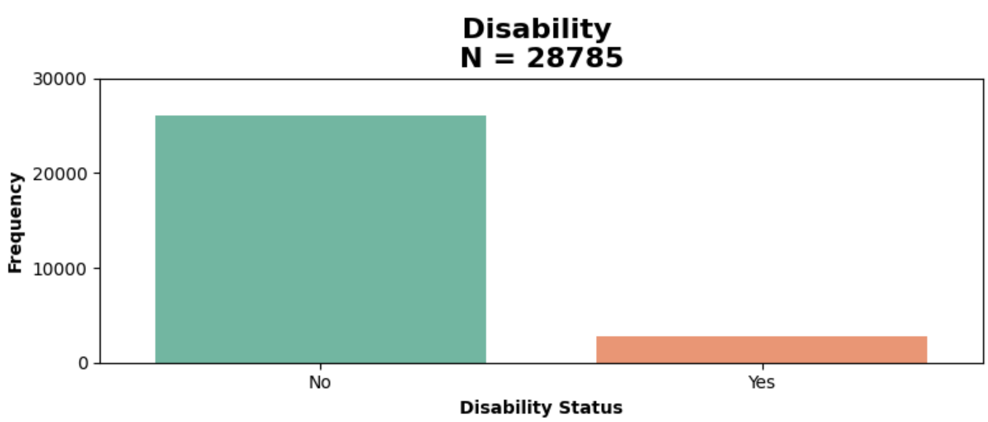


Figure 3: Histogram and Bar Charts of IMD, Region and Disability Distributions







Top 5 Enrolled Modules

Table 2: Top 5 Enrolled Modules
N = 31,284

Module Code	Frequency
BBB	7,692
FFF	7,397
DDD	5,848
CCC	4,251
EEE	2,859

Average Score by Module

To calculate module average scores, we use *weighted_score* from `studentassessment`. I joined *code_module, code_presentation* and *assessment_type* from `assessments` to `studentassessment` on *id_assessment*. However, exams are weighted at 100%, while TMAs and CMAs total 100% for each module's presentation period (Kuzilek et al., 2017). Combining these weighted scores results in a total equivalent to 200% of the module. Exams are thus listed separately, with only modules *CCC* and *DDD* having exams.

Table 3: Mean Score by Module, Averaging Over Presentation Periods and Assessments.

(a) TMA & CMA

Module Code	Average Score (Weighted)
EEE	80.0
FFF	74.4
CCC	73.7
BBB	71.5
AAA	69.0
DDD	68.7
GGG	0.0

(b) Exams

Module Code	Average Score (Weighted)
CCC	69.1
DDD	63.4

Top 5 modules with the Most Number of Fails

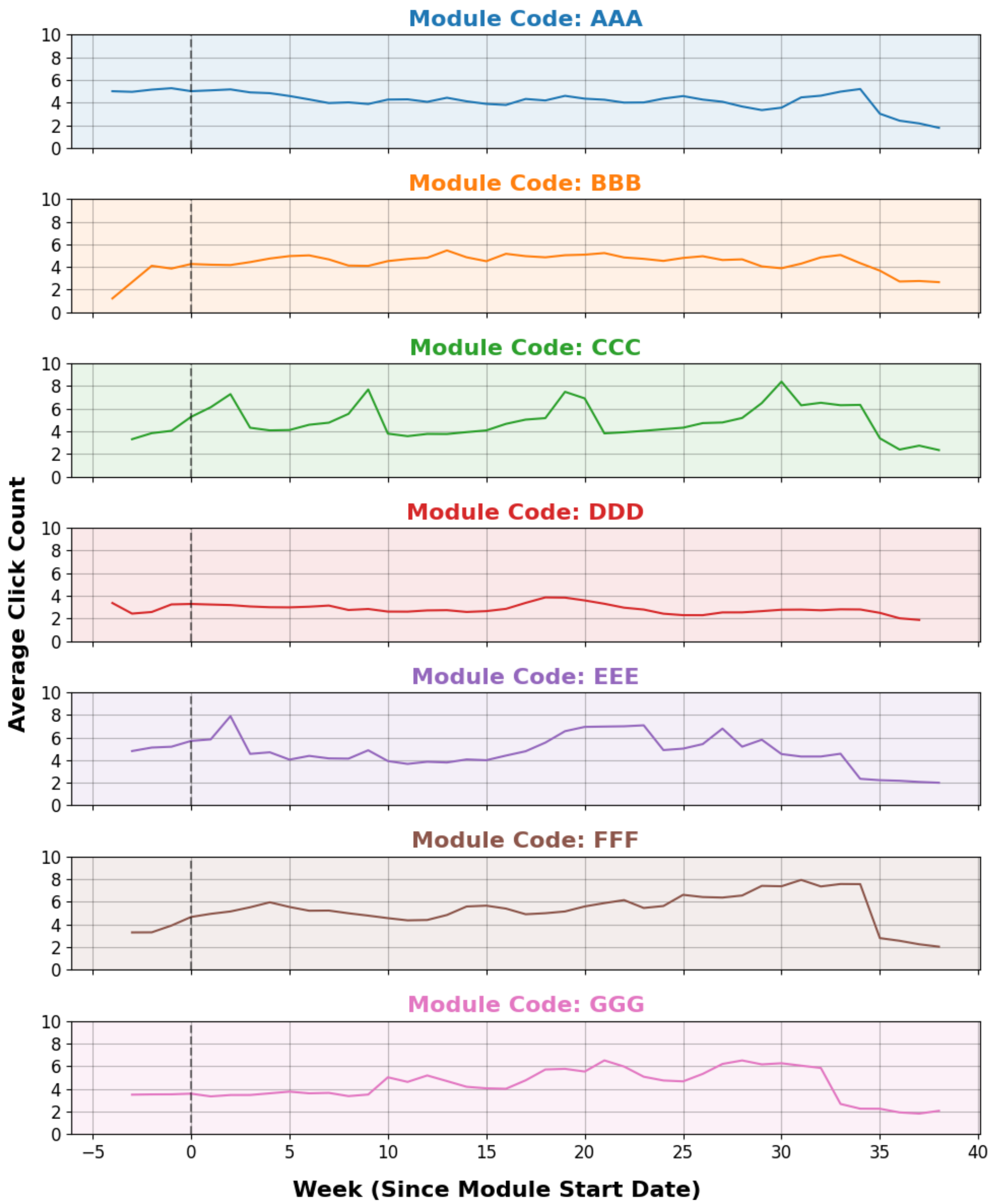
Table 4: Top 5 Modules with the Most Fails
N = 32,593

Module Code	Number of Fails
BBB	1767
FFF	1711
DDD	1412
CCC	781
GGG	728

Weekly Student Interaction with VLE

Figure 4 shows patterns in weekly VLE engagement across modules. Module *CCC* displays peaks in click counts, corresponding with CMAs (weeks 2, 9, 19, 30). Module *DDD* has consistently low clicks, suggesting the module may not utilise the VLE as much as others. All modules show a slight drop after week 35, potentially indicating late submissions.

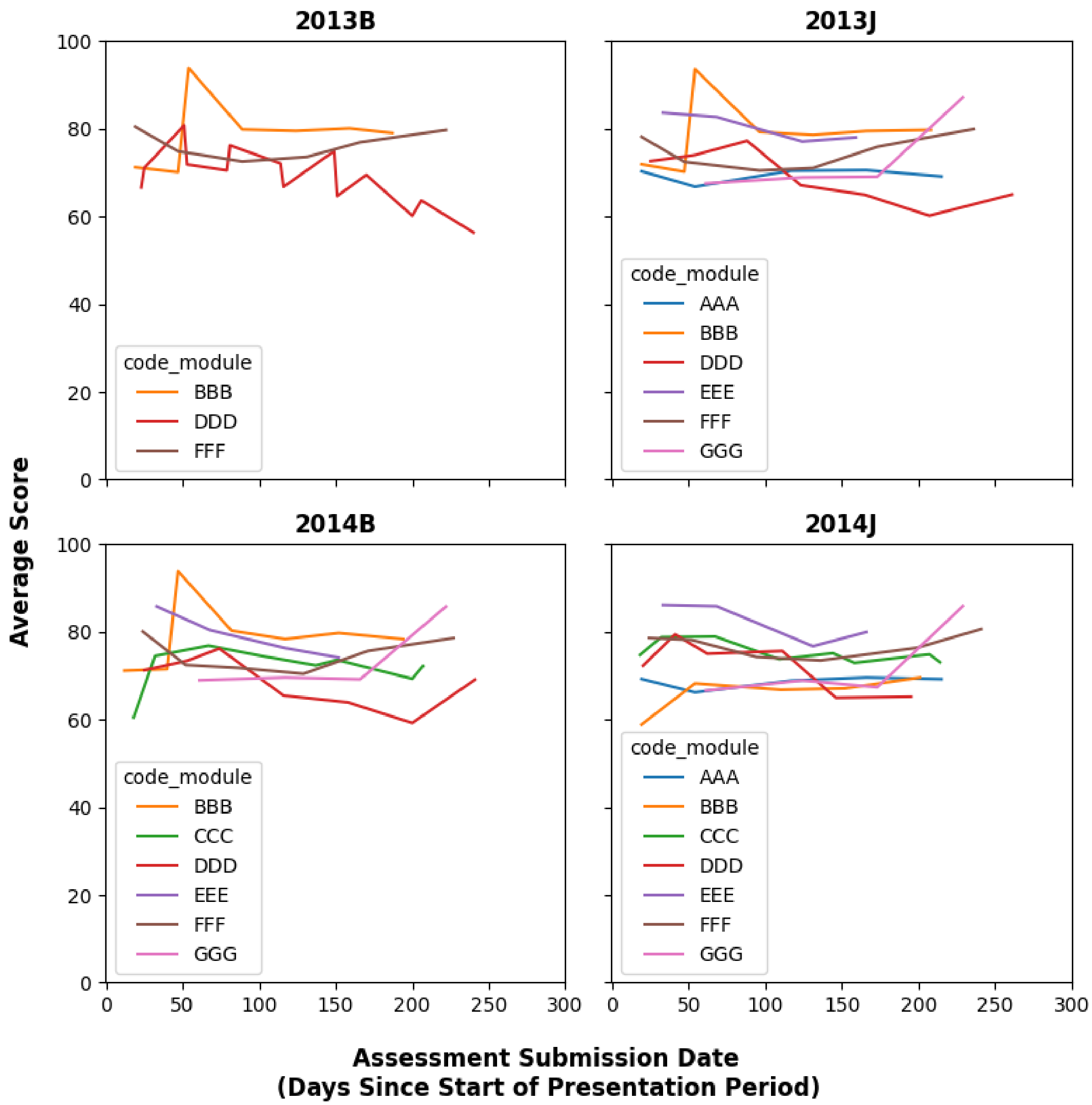
Figure 4: Line Graph Indicating Weekly Student Interaction with VLE Materials, by Module



Average Score Over Time

Figure 5 shows average scores remain stable across most modules, with module *DDD* showing a slight decline as it progresses, possibly indicating difficulty or a unique assessment method. By contrast, module *GGG* sees a notable increase in average score in its final assessment.

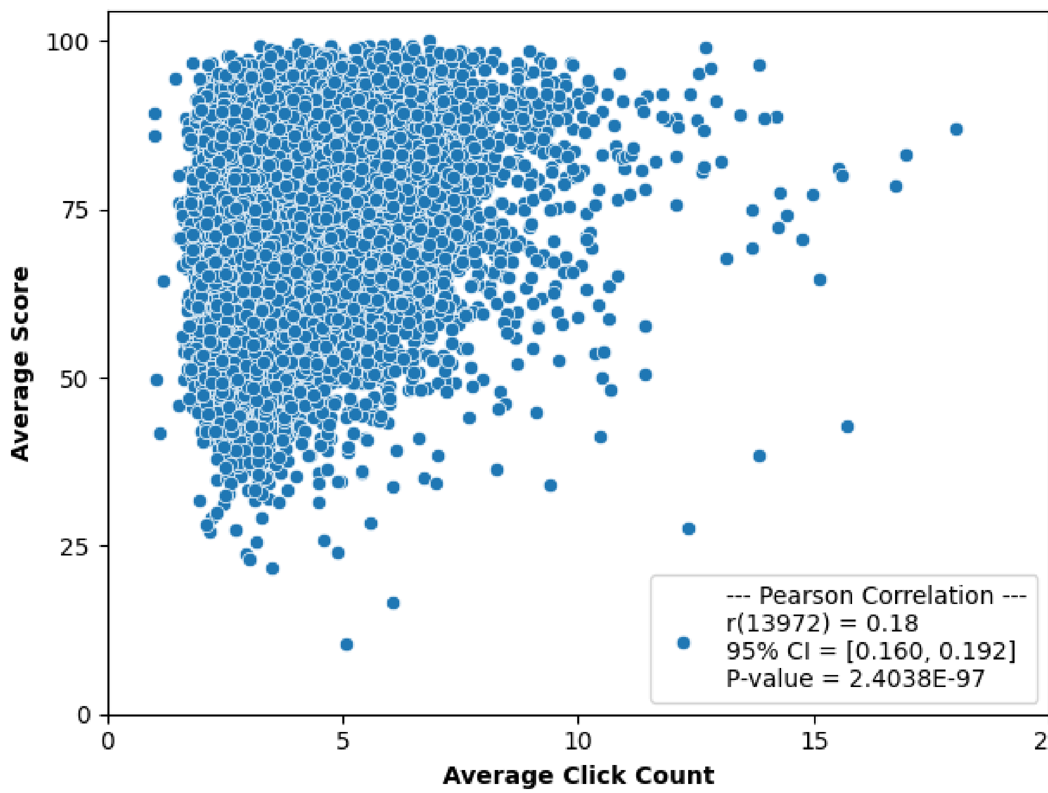
Figure 5: Line Graph Indicating Student's Average Scores on Each Module Assessment in Chronological Order, by Presentation Period



Correlation: VLE Engagement – Score

Figure 6 shows the correlation between students' average scores and their average click count. Although positive and significant at $\alpha = 0.05$, the correlation coefficient is small, $r(13,972) = 0.18, p = .24 \times 10^{-96}$.

Figure 6: Scatter Plot Showing the Correlation Between Students' Average Score and The Average Number of Clicks They Recorded on VLE Material



Despite a weak correlation, the direction remains positive, and **Figure 6** shows the variation in average scores slightly narrows as click count increases. Therefore, click count may still be relevant for measuring VLE engagement.

As discussed, we can measure VLE engagement by the number of unique days each student interacted with at least one VLE material. The correlation between this and students' average score in **Figure 7** shows a positive, though negligible, correlation, $r(13,972) = 0.21, p = .58 \times 10^{-141}$.

Figure 7: Scatter Plot Showing the Correlation Between Students' Average Score and The Number of Unique Days Spent Engaged with VLE Materials

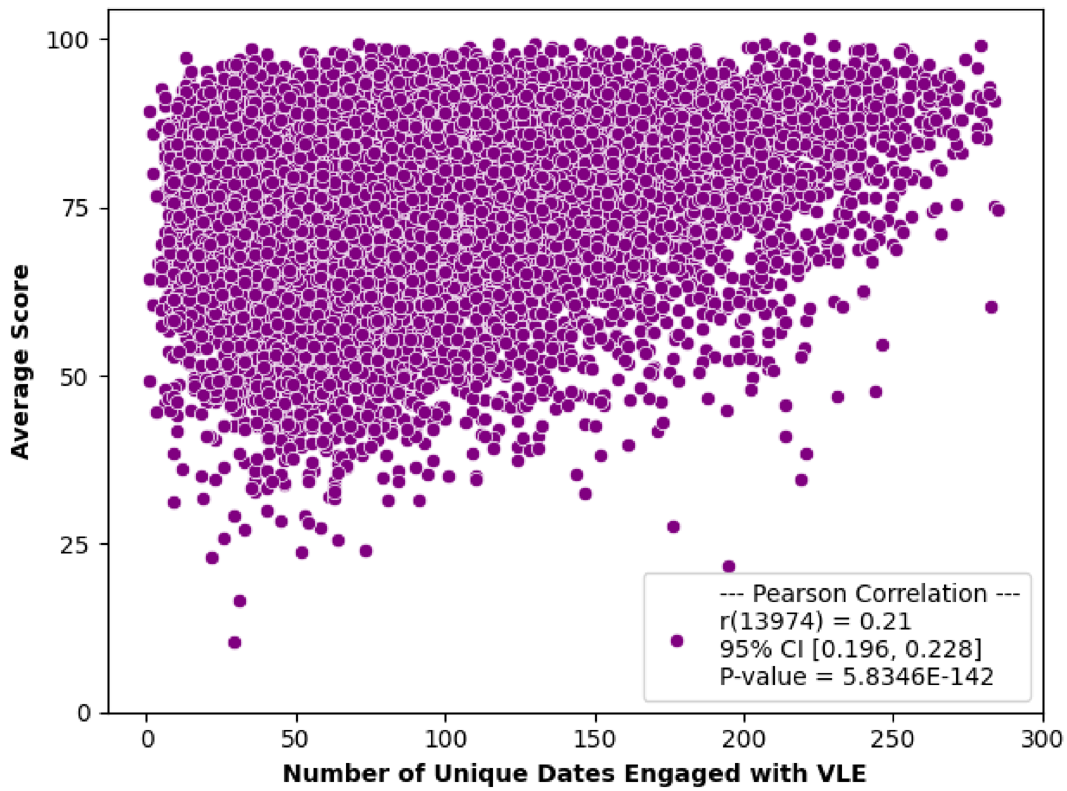


Figure 7 shows a more linear relationship than **Figure 6**. It also reveals that as student engagement with VLE materials on more unique days increases, the variation in average scores declines. These observations suggest that the number of unique days engaged with the VLE better captures the relationship between VLE engagement and students' scores. These two VLE engagement measures will be used together, with the number of unique days likely dominating in a composite measure due to its stronger, more linear relationship with students' average scores.

Methods

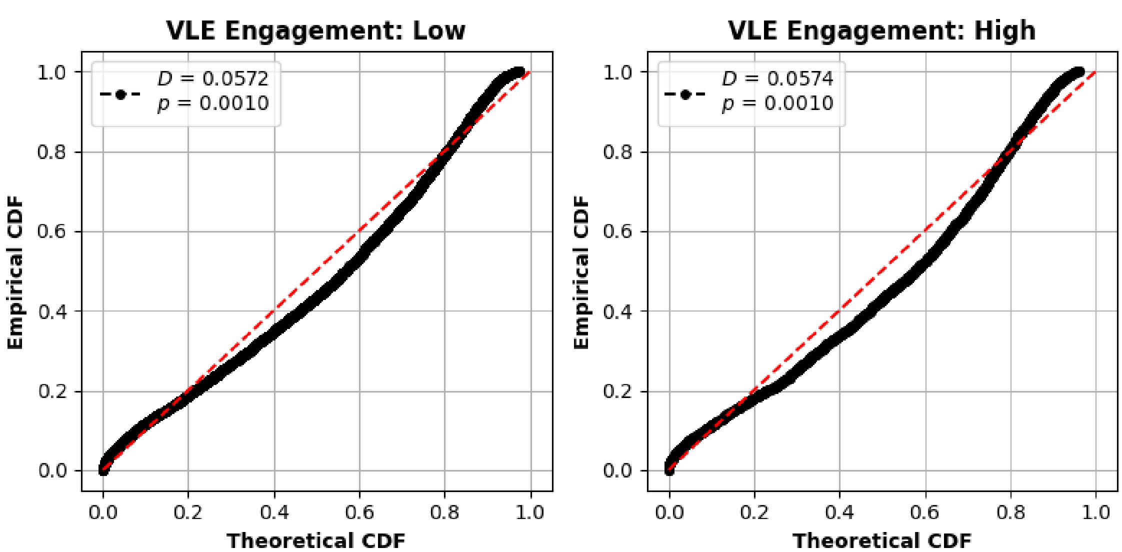
Is the VLE Improving Students' Grades?

Hypothesis Test

To test the hypotheses outlined in the introduction, a one-tailed Welch's t-test will assess whether the average assessment score of "High" VLE engagement students is significantly higher than that of "Low" VLE engagement students, with $\alpha = 0.05$ as the significance threshold. Such a hypothesis test assumes:

- Normality - that student score distributions at each VLE engagement level are roughly Gaussian. This can be checked with a p-p plot and Lilliefors tests ($\alpha = 0.05$). Although Figure 8 shows significant deviations from a Gaussian distribution, the large sample sizes for "Low" (N = 2,644) and "High" (N = 11,330) engagement suggest that the sample mean will approximate a Gaussian distribution, allowing for approximate normality. Therefore, Welch's t-test remains valid despite the normality violation.
- Independence - Each student is assigned to either the "Low" or "High" VLE engagement group based on their unique student ID.
- Unequal population variance - population variances do not need to be equal for Welch's t-test.

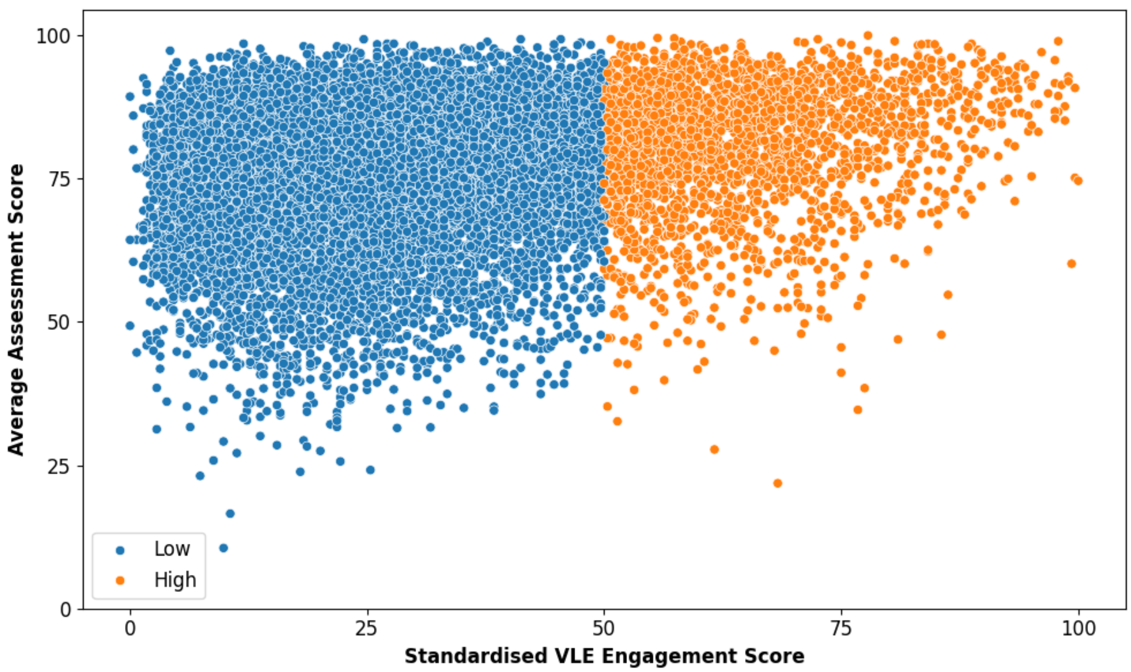
Figure 8: P-P Plot Comparing "Low" and "High" VLE Engagement CDFs to Gaussian CDFs



Data Pre-Processing

VLE engagement was measured using a composite score based on students' average click counts and the number of unique days they interacted with VLE materials. Principal Component Analysis (PCA) was used to create this composite measure. The principle component was scaled using min-max scaling to standardise VLE engagement scores between 0 and 100. Scores below 50 were classified as "Low", and scores ≥ 50 as "High". **Figure 9** shows the standardised VLE engagement measure against students' average scores, with hue indicating "Low" and "High" engagement.

Figure 9: Scatter Plot Comparing Students' Average Assessment Scores Against Their Standardised VLE Engagement Scores



Dataset

The dataset used includes 13,974 unique student IDs, each with an average assessment score, standardised VLE engagement score, and associated VLE engagement level. **Table 5** outlines summary statistics:

Table 5: One-tailed Welch's t-test Dataset Summary Statistics

		Score				**VLE Engagement**			
VLE Level	N	μ	σ	Min	Max	μ	σ	Min	Max
Low	11,330	75.64	12.11	10.56	99.33	24.68	12.47	0.00	49.9997
High	2,644	81.01	10.78	21.86	100.00	63.97	11.09	50.0001	100.00

Can we Predict Students' Grades?

To address this research question, a statistical model can be constructed to predict whether a student will pass their module, and examining how various factors relate to the probability of passing could offer insights to improve students' experience and module/VLE engagement.

Logistic Regression

With linear regression, the aim is to estimate predictor coefficients which minimise a loss function, such as the sum of squared residuals, enabling the value of the outcome variable to be predicted based on some set of predictor values. The outcome for linear regression is continuous, like height or exam score.

By contrast, logistic regression deals with a categorical outcome, such as whether a student passes (1) or fails (0). The coefficients in this model represent changes in the logarithmic odds of passing, with all other predictors constant, due to a unit increase in the predictor, rather than the direct change in the value of the outcome variable. In linear combination, these predictors yield a *logistic unit* (logit) of the outcome variable. To interpret this as a probability of being "true" (e.g. passing the module), the logistic function can be used:

$$p(1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

Our linear model is represented by the exponent of e, where β_0 is the intercept (logit when predictors are 0), each β_n is a model coefficient, and each x_n is the predictor value for a given observation. The exponent alone gives a logit, while the logistic function calculates the probability of an outcome variable being 1, given observed predictor values.

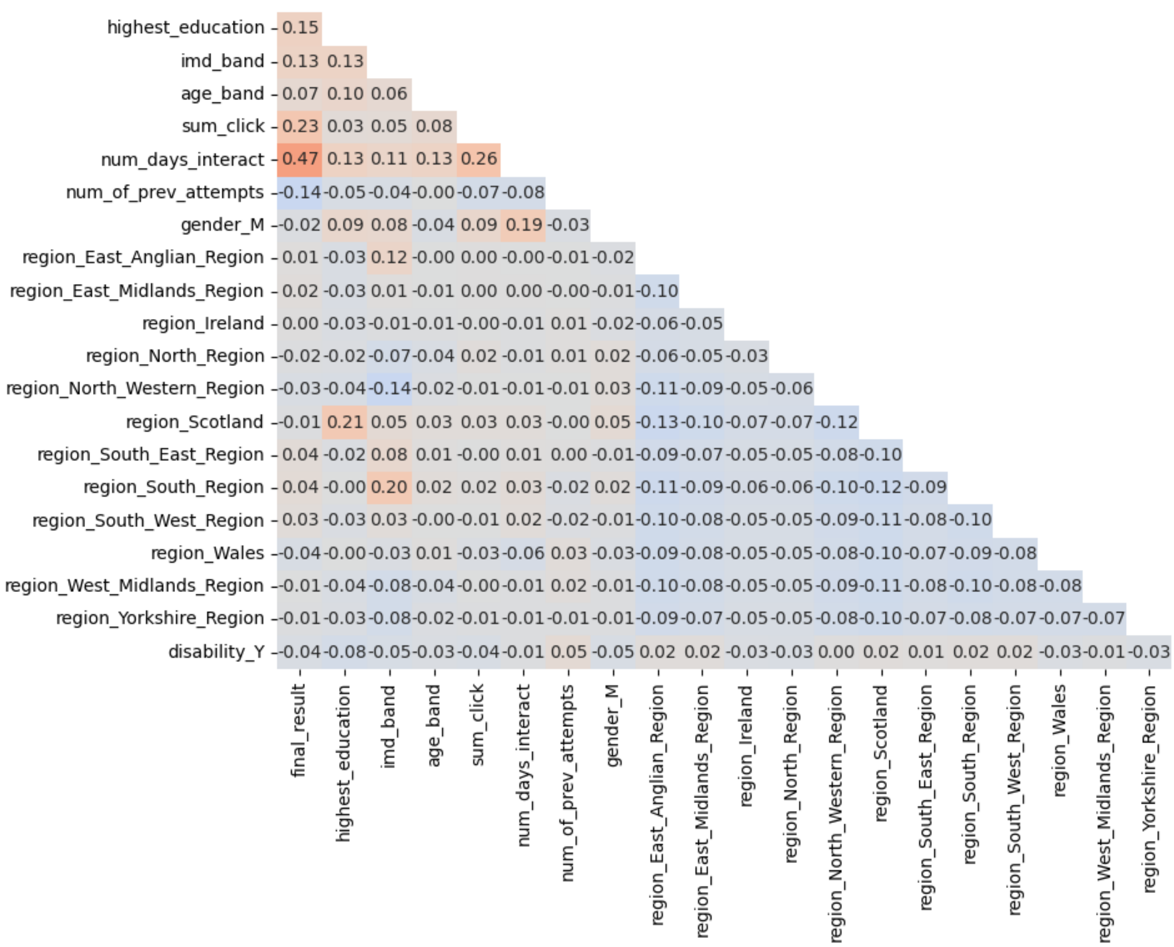
For example, if a logistic regression model predicts rain or no rain based on today's temperature and we have a predictor coefficient of -0.5 and an intercept of 3, the predicted probability of rain tomorrow, with a temperature of 12°C, would be:

$$p(rain|data) = \frac{1}{1 + e^{-(3-0.5(12))}} = 0.047\%$$

Assumptions

- Independence - Each observation in the data set should represent a unique module attempt. The *num_of_prev_attempts* feature indicates how many times a student has retaken a module, ensuring that repeated attempts are treated as distinct observations, and thus independence.
- Linearity - Using a logistic regression model assumes the relationship between each predictor and the log-odds of the outcome variable is linear.
- Multicollinearity - Predictors in logistic regression models shouldn't be highly correlated, which would reduce accuracy in estimating coefficients. Combining or excluding predictors can address this. **Figure 10** shows no significant correlations, with the highest being $r = 0.26$ between *sum_click* and *num_days_interact*, below the $|.3|$ threshold for practical significance.

Figure 10: Correlation Matrix for Features in Logistic Regression Model



Data Pre-Processing

The dataset for the regression model includes 20 predictors, 1 student ID (excluded from model), and 1 outcome variable (*final_result*). *final_result* was encoded as 0 for "fail", 1 for "pass" or "distinction", with 10,156 "withdrawn" rows removed.

Regarding predictors, 3 were continuous (*sum_click, num_days_interact & num_of_prev_attempts*), 3 ordinal with evenly-spaced encodings between 0 and 1 (*highest_education, imd_band, age_band*) and the remaining 14 were nominal with binary encodings (0 for "false", 1 for "true") (*gender_M (i.e. whether or not the student was male), region_East_Anglian_Region, region_East_Midlands_Region, region_Ireland, region_North_Region, region_North_Western_Region, region_Scotland, region_South_East_Region, region_South_Region, region_South_West_Region, region_Wales, region_West_Midlands_Region, region_Yorkshire_Region, disability_Y (i.e. whether or not the student had a disability)*).

Since many features are nominal, a reference category for each of these is implicit in this model. For gender, an observation with 0 in *gender_M* is therefore female, 0 in *disability_Y* refers to no disability, and 0 in every region feature means the student is from London.

Lastly, there was a large class imbalance between "fail", N = 8,302 (32.28%), and "pass" rows, N = 17,416 (67.72%). To use our regression model as a classifier, having this discrepancy will disproportionately bias the classifier towards placing predictions in the larger class (i.e. "pass"), inflating the false positive rate (i.e. reducing the specificity of the model). To address this, given the smallest class size is still large, a random sample of 8,302 rows was drawn from the "pass" class, excluding 9,114. As such, both classes were of size N = 8,302 prior to model construction.

Train-Test Procedure

Once created, the model's performance on unseen data should be assessed. To achieve this, we can shuffle then split the data set into two subsets. The first subset contains 80% of the data and is used to "train" the logistic regression model. The remaining 20% is therefore "unseen" by the model and so can be used to "test" how well the model does at correctly classifying unseen data.

The results of the "test" phase can be shown in a confusion matrix - a 2x2 matrix comparing the model's predictions of which class an observation in the test sample belongs to with the actual class it belongs to. Ideally, a perfect model would have 0 in both the actual-fail to predicted-pass and actual-pass to predicted-fail sections, and all the model predictions should instead fall in the sections where the prediction matches the actual class (true positive and true negatives). In reality, the model will likely make mistakes, resulting in false positives and false negatives.

With this confusion matrix, a handful of metrics can be used to assess model performance. Firstly, accuracy will tell us the percentage of classifications made by the model that were correct. Note that for a logistic regression model, a random classifier would achieve a long-run accuracy of 50%, and so our model should at least need an accuracy higher than 50%. Precision tells us the ratio of true positives to total predicted positives. Recall is the ratio of true positives to the actual number of positives, whilst specificity is similar to recall but applies to true negatives and actual negatives. For these three, a value closer to 1 is desired.

Two further metrics help to provide a more overall summary of the model. The F1 score combines precision and recall to provide a score from 0 to 1, indicating the model's ability to classify true positives whilst minimising the number of false negatives and positives. As such, the F1 score is concerned with the model's performance on the positive class (predicting "passes"), informedness combines recall and specificity to assess the performance on both the positive and negative class - again, a score closer to 1 is desired.

Lastly, precision-recall (PR) and receiver operating characteristic (ROC) can be used to assess the model's performance at different probability thresholds for classification than just the 0.5 used up to now.

Feature Selection

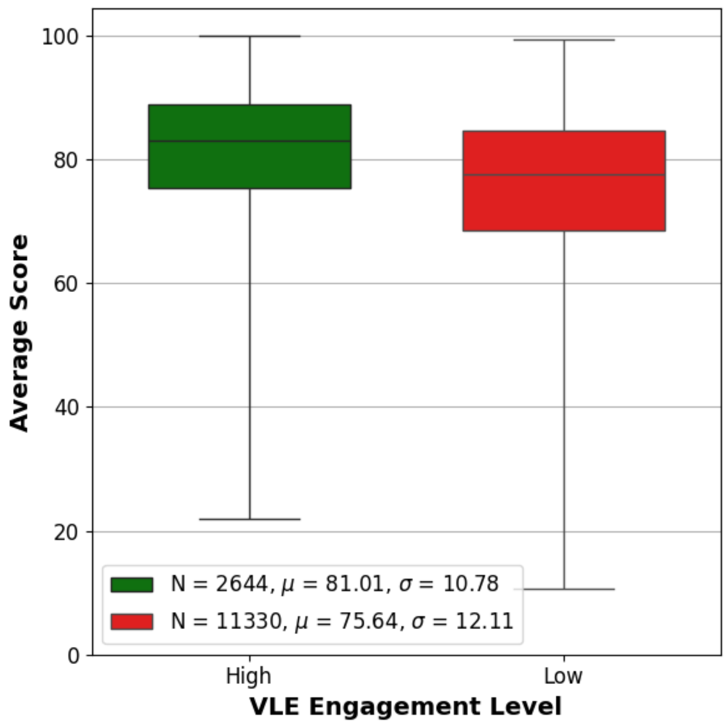
Initial feature selection involved choosing features from the data set which described particular characteristics about each student. These features were outlined earlier but can be broadly categorised as either referring to demographic information or to module engagement. Whilst this selection could be refined further using techniques such as recursive feature elimination, given that no features were highly collinear (see **Figure 10**) and that we are aiming to both predict final results and explore the relationship between final results and the feature set, I have decided to not reduce the feature set.

Results

Is the VLE Improving Students' Grades?

Results from the Welch's t-test indicate that "High" VLE engagement students have a significantly greater mean average assessment score compared to "Low" VLE engagement students, $t(13{,}972) = 22.51, p = 1.74 \times 10^{-106}$. **Figure 11** compares each groups' distributions.

Figure 11: Boxplot Comparing Score Distributions of "High" and "Low" VLE Engagement



Whilst the difference is statistically significant, this may be due to the large sample sizes of both groups ($N_{high} = 2{,}644, N_{low} = 11{,}330$), wherein a two-sample t-test is more likely to find a statistically significant difference. As such, whilst we may be confident in claiming "High" VLE engagement average assessment scores are significantly higher than "Low" VLE engagement, we can see that the actual value for these average assessment scores are fairly close to each other ($\mu_{high} = 81.01, \mu_{low} = 75.64$).

Can We Predict Students' Grades?

Logistic Regression Model

Our regression model can be expressed in the form:

$$y = b_0 + \beta_1 highest\_education + \beta_2 imd\_band + ... + \beta_{19} region\_Yorkshire\_Region + \beta_{20} disability\_Y$$

Applying the model to the logistic function yields our function for estimating the probability of a given observation passing:

$$p(Pass|Data) = \frac{1}{1 + e^{-b_0 + \beta_1 highest\_education + \beta_2 imd\_band + ... + \beta_{19} region\_Yorkshire\_Region + \beta_{20} disability\_Y}}$$

Regression Results

Table 6 outlines the results of the logistic regression. Our intercept coefficient (-2.8090) refers to the log-odds of passing when all predictors equal 0, equivalent to a probability of 5.68%. In other words, our model's reference category is a 0-35 year-old female student with no disabilities, in the lowest IMD band with no VLE engagement from London; this person has a probability of passing of 5.68%.

highest_education, imd_band, sum_click and *num_days_interact* have significant positive relationships to the log-odds of passing, with a unit increase in *highest_education* equating to the largest increase in the log-odds (1.4856) of passing. *region_East_Midlands_Region, region_South_East_Region, region_South_Region* and *region_South_West_Region* also share a positive relationship with the outcome variable, suggesting students from these regions have increased log-odds of passing relative to those from London.

num_of_prev_attempts, gender_M, region_Scotland and *disability_Y* all share negative relationships with the outcome variable, suggesting retaking modules, being male, living in Scotland or being someone with a disability reduces one's log-odds of passing.

Lastly, *age_band, region_East_Anglian_Region, region_Ireland, region_North_Region, region_North_Western_Region, region_Wales, region_West_Midlands_Region* and *region_Yorkshire_Region* are all non-significant predictors.

Table 6: Results: Logit

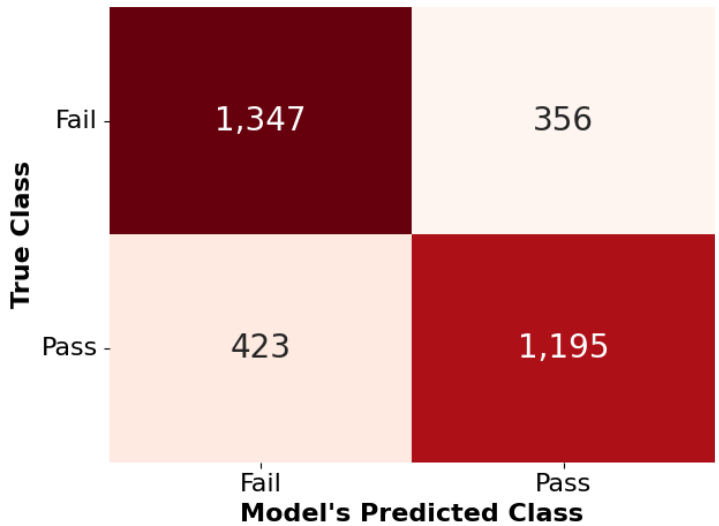
Model:	Logit	Method:	MLE
Dependent Variable:	final_result	Pseudo R-squared:	0.281
Date:	2025-01-03 13:38	AIC:	16596.0673
No. Observations:	16604	BIC:	16758.1327
Df Model:	20	Log-Likelihood:	-8277.0
Df Residuals:	16583	LL-Null:	-11509.
Converged:	1.0000	LLR p-value:	0.0000
No. Iterations:	6.0000	Scale:	1.0000

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
Intercept	-2.8090	0.0972	-28.9100	0.0000	-2.9995	-2.6186
highest_education	1.4856	0.1101	13.4988	0.0000	1.2699	1.7013
imd_band	0.5375	0.0657	8.1812	0.0000	0.4088	0.6663
age_band	-0.0111	0.0842	-0.1322	0.8948	-0.1761	0.1538
sum_click	0.1802	0.0122	14.7758	0.0000	0.1563	0.2041
num_days_interact	0.0250	0.0005	53.4641	0.0000	0.0241	0.0259
num_of_prev_attempts	-0.6400	0.0445	-14.3935	0.0000	-0.7272	-0.5529
gender_M	-0.8379	0.0409	-20.4700	0.0000	-0.9182	-0.7577
region_East_Anglian_Region	0.0951	0.0854	1.1138	0.2654	-0.0723	0.2625
region_East_Midlands_Region	0.2751	0.0936	2.9386	0.0033	0.0916	0.4586
region_Ireland	0.2338	0.1242	1.8822	0.0598	-0.0097	0.4773
region_North_Region	0.0849	0.1207	0.7037	0.4816	-0.1516	0.3215
region_North_Western_Region	-0.0227	0.0878	-0.2580	0.7964	-0.1947	0.1494
region_Scotland	-0.2920	0.0835	-3.4975	0.0005	-0.4557	-0.1284
region_South_East_Region	0.3897	0.0980	3.9750	0.0001	0.1975	0.5818
region_South_Region	0.1796	0.0902	1.9918	0.0464	0.0029	0.3564
region_South_West_Region	0.2754	0.0933	2.9511	0.0032	0.0925	0.4583
region_Wales	-0.0063	0.0932	-0.0676	0.9461	-0.1889	0.1763
region_West_Midlands_Region	0.1416	0.0910	1.5563	0.1196	-0.0367	0.3199
region_Yorkshire_Region	0.1200	0.0959	1.2522	0.2105	-0.0678	0.3079
disability_Y	-0.2654	0.0681	-3.8966	0.0001	-0.3989	-0.1319

Model Performance

Using the same feature set, an 80-20 train-test split was performed. **Figure 12** shows the confusion matrix for the model's predictions.

Figure 12: Confusion Matrix for Logistic Regression Model Predictions



Given the data set was balance prior to model construction, it is reasonable to observe false positive and false negative rates that are similar. K-fold cross validation with 10 subsamples provided a mean accuracy score of 0.76 ± 0.01. **Table 7** outlines the model's performance metrics.

Table 7: Model Performance Metrics

Accuracy	0.78
Precision	0.78
Recall	0.76
Specificity	0.79
F1 Score	0.77
Informedness	0.55

These metrics suggest the model has performed reasonably well, albeit informedness is somewhat low. Moreover, **Figures 13** and **14** display good performance on PR AUC (0.71) and ROC AUC (0.78)

Figure 13: Precision Recall Plot for Logistic Regression Model

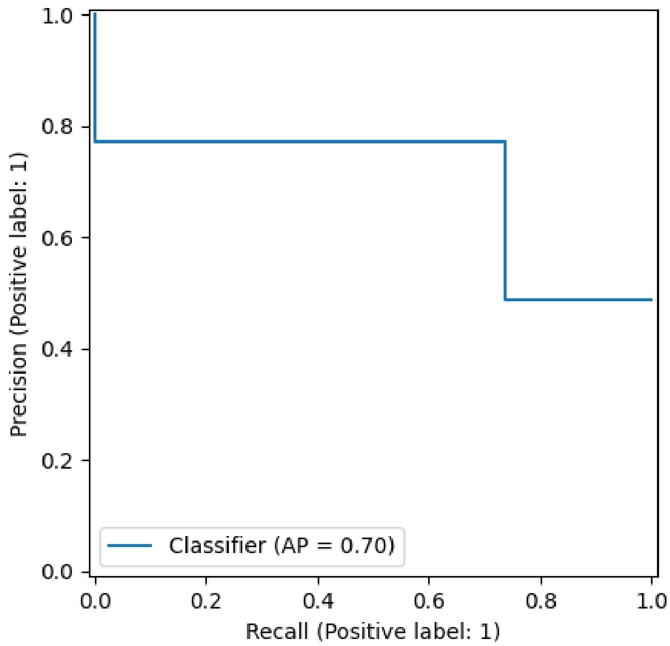
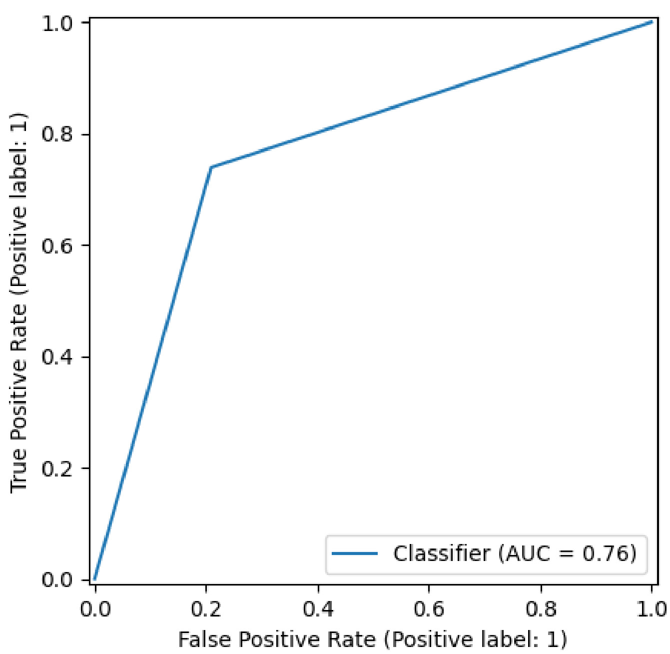


Figure 14: ROC Curve Plot for Logistic Regression Model Predictions



Discussion

Is the VLE Improving Students' Grades?

In determining whether the VLE had improved student's grades, a composite measure of VLE engagement was created using average click count and number of unique days engaged with VLE material. Splitting students into "High" and "Low" VLE engagement groups, a one-way Welch's t-test found "High" VLE engagement students had significantly higher average assessment scores compared to "Low" VLE engagement. As such, this finding provides evidence to suggest that the VLE has improved students' grades as we would expect to see this pattern if VLE engagement did actually improve grades.

However, there could be alternative factors which lead to the difference in average score between these two VLE engagement groups. Interestingly, the logistic regression model does provide indication of a positive relationship between VLE engagement and average score, as *sum_click* and *num_days_interact* share a positive relationship with the probability of passing. Given that passing a module necessarily means a higher average score than failing, this finding does help to support the view that VLE engagement does improve grades.

To address this question more thoroughly, an experimental design would be preferable. Given the ethical issues with denying certain students access to the VLE for weighted assessments, an experiment could be used with unweighted assessments comparing a randomised control group with no VLE engagement against a random sample of students who are given VLE access. Those students who engage more so with the VLE can have their grade on the unweighted assessment compared to those who had no VLE access. Such a design would enable potential confounds to be controlled for, such as extracurricular commitments, which might otherwise be the actual causal factors being observed in the present analysis.

Limitations aside, the results observed are in line with the alternate hypothesis and so increased encouragement of VLE usage among students may lead to an increase in average assessment scores. It should be noted that the actual difference in scores is small (81.01 to 75.64), suggesting that if the VLE does improve grades the effect size is possibly negligible. As such, whilst there is evidence to suggest VLE engagement improves grades, it may not be an efficient mechanism for improving student grades compared to alternative options. It is recommended that VLE engagement be encouraged to those who currently do not utilise it much, but for those who already do, alternative means should be explored to improve grades.

Can We Predict Students' Grades?

In constructing a logistic regression model to predict whether or not a given module attempt would result in a pass, it was found that education level, IMD band and VLE engagement all were significant positive predictors, along with being from the East Midlands and the South (relative to London). Retaking modules, being male, from Scotland or having a disability all were significant negative predictors. The remaining predictors were non-significant.

The implication is that there is evidence of disparities in passing modules across a number of student demographics. Poorer students, those with lower education backgrounds, male students and those with disabilities are all significantly less likely to pass modules. Future research could focus on attempting to understand factors which contribute to why these demographics are less likely to pass and address them where possible. The same applies to the evidence of some regional disparities.

Model prediction metrics suggest the model is good at correctly classifying students, although with room to improve. Such improvement could come from considering additional measures of students' academic performance and engagement. At present there are just 2 VLE engagement measures. It would be beneficial to have additional information, such as which specific module materials were accessed and how often, rather than generic indicators of material type like "home page" and an overall click count. More detail on VLE engagement would also help to highlight exactly what about the VLE helps to improve student grades, rather than just seeing if the VLE as a whole improves grades.

It is recommended that open dialogue with disadvantaged demographics be pursued to determine potential barriers to success. This could be in the form of interviews or surveys to gauge sentiment, measure studying behaviours and identify risk factors among these demographics that could be investigated further. The model also provides a useful tool for identifying students at risk of failing and so could be incorporated into this process of reducing grade disparities for such demographics. Furthermore, it is evidenced from the model that VLE engagement shares a positive relationship with grades, further underlining the need to promote its usage to those who currently do not engage with it. Such promotion could come in the form of workshops showing the benefits of the VLE or raising awareness of the evidence presented in this report highlighting the positive relationship between VLE engagement and grades.

Conclusion

In conclusion, this report investigated whether the VLE improved students' grades and constructed a model to predict whether or not a given student will pass their module. Whilst the influence of the VLE on grades may be small, it is evident from both the hypothesis test and results from the regression model that encouraging VLE engagement among "Low" engagement students will help to improve grades. This could be achieved through awareness campaigns or workshops. Furthermore, it is recommended that further investigation be undertaken to determine barriers to success for student groups who are currently less likely to pass modules, according to the regression model. Engaging with groups who are currently at greater risk of failing to understand potential risk factors through surveys, interviews and further research is recommended.

References

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. Scientific Data, 4(1), 170171. https://doi.org/10.1038/sdata.2017.171

Mosley, N. (2024, November 12). 2023 Review: What's the state of the VLE market in UK higher education? Neil Mosley Consulting. https://www.neilmosley.com/blog/2023-review-whats-the-state-of-the-vle-market-in-uk-higher-education

Appendicies

Appendix A: Frequency of *id_site* and *id_student* Amongst *sum_click* Counts > 1,000

Table 8: Frequency of *id_site* and *id_student* Amongst *sum_click* Counts > 1,000

id_site	Frequency
526721	1
526853	1
546703	1
551035	1
551135	1
673519	1
716238	1
716434	1
716831	1
729798	1
729809	1
729813	1
729815	1
729844	1
773028	1
790888	1
814061	1
832729	1
909032	1
909096	1
909298	1
909314	1
909315	1
913490	1
913671	1

id_student	Frequency
582087	3
306202	2
285170	2
649840	2
204505	1
491011	1
368315	1
498031	1
497180	1
592214	1
601224	1
605868	1
543643	1
618606	1
620337	1
633902	1
678395	1
687378	1
687563	1
1894188	1
2616950	1