Factors Influencing Success in Online College

MGT 6203: Data Analytics in Business

Team GitHub | OULAD Dataset

Team 111: Brian McClure, Richard Han, Maysam Molan, John Redden

Background: One of the major areas of focus for educational administrators and faculty is the drop, fail, withdraw, and incomplete rate (DFWI) rate. This metric is essential when creating strategies to support struggling students and improve student outcomes; a high DFWI rate means that a program or course suffers from student attrition. Thus, an institutional goal is to reduce the DFWI rate efficiently and economically while maintaining quality programs. To this end, data analytics can play a definitive role in identifying which student demographic features, course activity, and assessment patterns influence this rate.

The insights from this project are designed to inform business decisions within the educational sector. By identifying "at-risk" students and significant factors contributing to their challenges, our models will facilitate the development of more informed intervention strategies. Many studies of this type have identified significant factors such as prior academic achievement, student demographics, e-learning activity, psychological attributes, and learning environments (Alyahyan, 2020). The Open University Learning Analytics Dataset (OULAD) has many of these factors, which gives us confidence that we are on the right track.

Research Questions and Hypotheses: Informed by the Alyahyan study, we hypothesize that the DFWI rate is significantly influenced by four primary factors: student demographics, prior academic performance, course engagement metrics, and assessment performance. Second, we propose that each factor independently affects the failure and withdrawal rates. Third, we hypothesize that incorporating assessment and online engagement data, in addition to student demographics data, will improve predictive performance of our models. Our approach in this paper focuses on answering three distinct and related questions:

1. What predictors affect student outcomes measured by DFWI rate?
2. What early indicators can signal that a student may be at risk of failing or withdrawing?
3. Can we predict student success in a timely manner for actionable intervention?

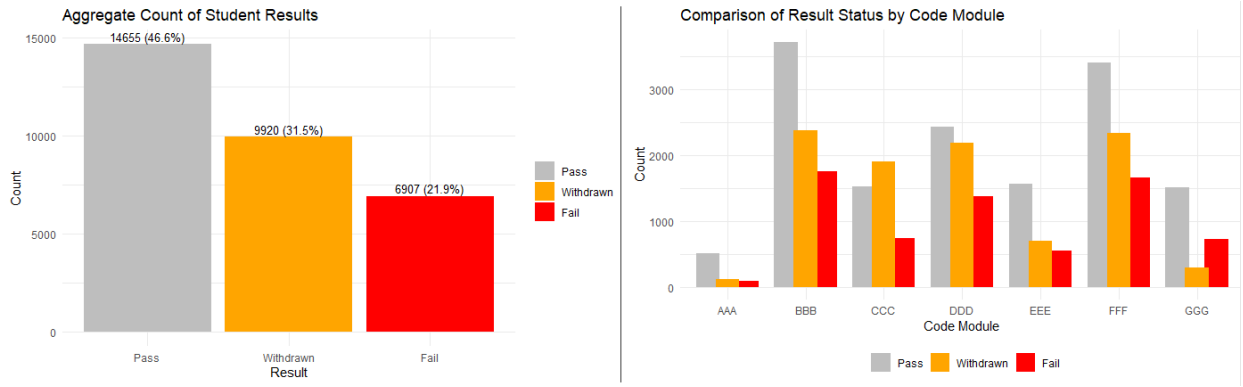
To answer these questions, we employed logistic regression to determine significant features, analysis of regional deprivation, 7-day moving averages of daily click data, CUSUM analysis of assessments, and feature engineering leading to a comprehensive predictive model.

Business Justification: Our study employs classification, statistical analysis, and time series modeling to develop targeted intervention strategies for at-risk student populations. Furthermore, we offer ***actionable interventions*** by emphasizing true "Data Driven" decisions. Such analytics are often highlighted but rarely implemented in education (Custer et al., 2018); our analysis can provide significant economic benefits to both students and businesses relying on a well-educated workforce.

This analysis aims to assist students, faculty, and administrators by offering early warning signals of at-risk students, facilitating timely interventions. Our findings and actionable interventions are presented to help educators enhance course designs with data-driven insights into critical course features and to aid administrators in boosting operational efficiency through increased enrollments and reduced DFWI rates. Our focus on model interpretability ensures that our findings and recommendations are accessible to all stakeholders, promoting informed decision-making throughout the educational ecosystem.

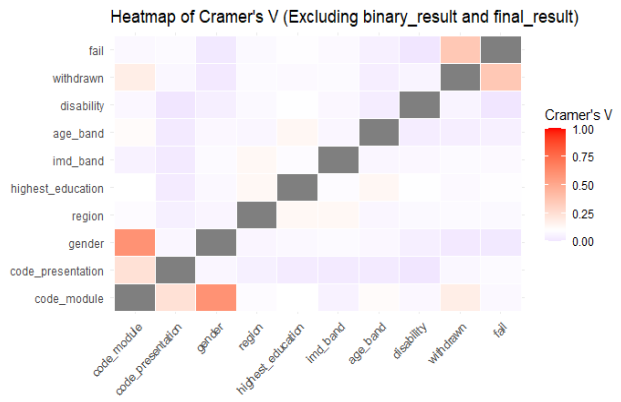
Overview of Data: The Open University Learning Analytics Dataset (OULAD) is a large, anonymized collection of data on courses, students, and their interactions within the Virtual Learning Environment (VLE). The dataset is comprised of 32,593 student records from 22 sections across seven different courses, spanning years 2013 to 2014 (Kuzilek, J., Hlosta, M., & Zdrahal, Z., 2017). It includes demographic data (e.g. gender, disability, age, socioeconomic band, etc.) and scholastic data (date of registration, number of credits earned, etc.), which can be used for modelling key variables affecting DFWI rate. Furthermore, there are two separate datasets that show student scores on their assessments and how much they interact with the digital learning environment (measured by the number of clicks across time). These datasets are cleaned and made into time series for each student.

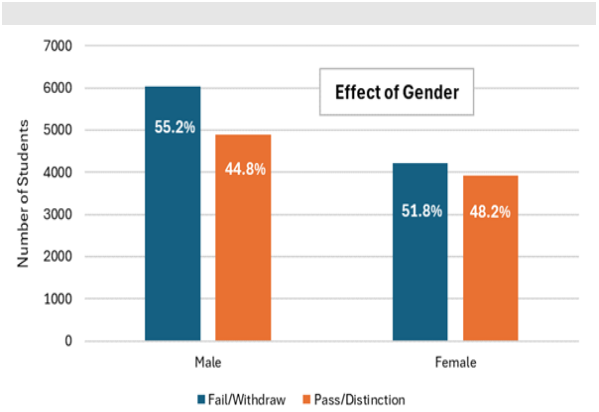
Exploratory Data Analysis (EDA): Out of all students in the dataset, in aggregate, about 46.6% of students pass/have distinction, 31.5% withdraw, and 21.9% fail. These facts can be further computed by courses which are labelled AAA through GGG.



The main objective is to analyze the significant factors affecting this 53.4% unsuccessful Withdrawn/Fail rate. To this end, we created two new binary variables `withdrawn` and `failed` which formed the target of two logistic regression models. After regressing on the binary `withdrawn` variable, at the 5% level of significance, we find that course (`code_module`) was significant as well as regional deprivation metrics (`imd_band` and `region`). Furthermore, confirming the Alyahyan study, we see that academic preparedness factors (`studied_credits` and `highest_education`) are also significant. Next, we regress on the new `failed` variable and found that the significant features to be very similar to those in the withdrawn regression. Surprisingly, for this failed cohort of students, `disability` was not significant as it was for the withdrawn cohort. Males tend to have a higher occurrence of failing and withdrawing compared with women shown in the figure below.

Because most of our features are categorical, we found that the following Cramer's V heatmap offers a visual representation of the strength of association between them. Here we are looking for feature associations between the new `withdrawn` and `fail` variables.





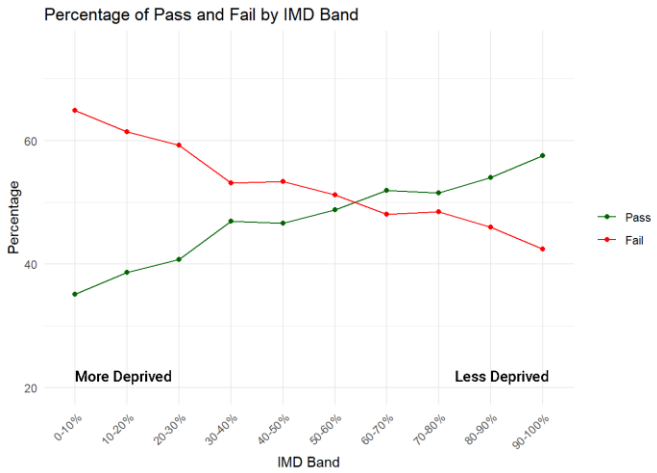
Notably, there's a high correlation between `code_module` and `gender`, as well as between `code_module` and `imd_band`, which suggests that both the gender of students and their socioeconomic background are linked to the specific modules they enroll in. Additionally, a moderate association is observed between `imd_band` and `withdrawn`, hinting that socioeconomic factors could influence fail and withdrawal rates. These insights were used to guide further analysis, focusing particularly on outcome patterns and how they relate to regional and economic backgrounds, aligning with our goal of reducing the DFWI rate.

In the end we found the following significant demographic features produced the best predictive models for a simple Pass/Fail outcome: `code_module + highest_education + imd_band + num_of_prev_attempts + studied_credits + disability`. We then use these features and cross validation to create three baseline classification models (Logistic Regression LR, Random Forest RF, and K-Nearest Neighbors KNN) that all use the stated demographic data. Based on initially calculated accuracies they all perform similarly.

	LR	**RF**	**KNN**
Student Demographics	62.1%	62.1%	63.2%

These initial results using the demographic data only are insufficient. Therefore, we explore methods that aim to improve these models by adding new interaction features from the click data and assessments data that may subsequently improve these results.

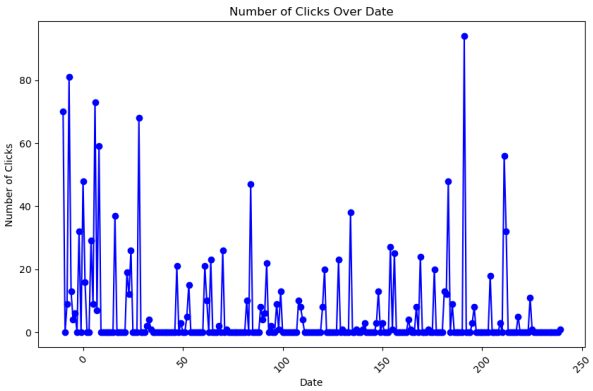
Study 1 (Poverty Band): The Index of Multiple Deprivation (IMD) is a multidimensional ***measure of poverty*** that considers various aspects of deprivation, including income, employment, health, education, crime, the living environment, and access to housing within an area.



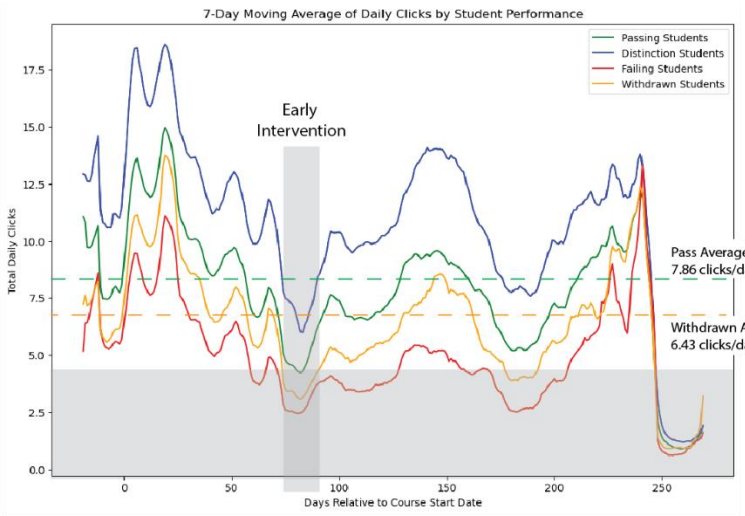
This index is crucial for identifying disparities in living conditions and is often used to allocate resources and target interventions in the most deprived areas. Since our EDA finds this metric to be quite significant in the DFWI rate, we graph two variables-- "Pass," which includes distinction, and "Fail," which includes withdrawals--over the provided deprivation categories. The lower IMD ranges signify higher levels of poverty (**more deprived**) whereas higher bands signify lower levels of poverty (**less deprived**). It is astonishing to visualize the relationship between pass and fail rates with respect to the regional deprivation metric. The graph reveals a clear pattern that highlights the disparity in educational outcomes in relation to varying levels of poverty. This trend is inverted at the crossing of the "Pass" and "Fail" lines in the midpoint of the IMD band and leads us to our first targeted intervention. To support students in more deprived areas with IMD less than 50%, we suggest focusing on efforts that reduce barriers. These interventions include additional tutoring, mentorship programs, and academic counseling. A goal is to work towards leveling the educational playing field and provide all students with an equal opportunity to succeed academically, regardless of their socioeconomic background.

Study 2 (Clickstream Time Series Data): For the virtual learning environment (VLE) analysis, the ***studentInfo.csv*** dataset and the ***studentVle.csv*** dataset are most relevant. We want to understand the relationship between student activity and the student outcome for each course module. The variable `sum_click` in the VLE dataset tells us how many clicks the student performed in a given course for each day. Aggregating the number of clicks by day gives us total daily clicks, which we can observe as a time series. For example, here is the clickstream data for a passing student in one of the courses.

As we can see, this student had a lot of consistent interaction with the virtual learning environment from beginning to end of the term. In contrast, the clickstream data for a typical failing student is sparse and shows less interaction. For a withdrawn student, their clickstream data may be cut short due to leaving the course early on.



Although the clickstream data we have observed so far corroborates our expectations for each type of student, we must be careful not to generalize since there are some failing students with high total daily clicks and some withdrawn students that persist for a long period of time. To get a general sense of how each type of student behaves, here are the 7-day moving averages for the mean total daily clicks for each type of student.



In general, the mean total daily clicks for distinction students are higher than those for passing students, which are higher than those for withdrawn students, which in turn are higher than those for failing students. The average total daily clicks for distinction, pass, withdrawn, and fail students are 10.7, 7.86, 6.43, and 4.97, respectively. The standard deviations for distinction, pass, withdrawn, and fail students are 4.70, 3.81, 3.42, 3.21, respectively. Passing and distinction students have overall more clicks than failing and withdrawn students; and, passing and distinction students have more variation in their clicks than failing and withdrawn students. Thus, the standard deviation and average number of clicks may play a role in distinguishing the different types of students. This leads us to our second actionable intervention. We can alert instructors and teaching assistants when a student's moving average of daily clicks across the first 75 days falls below 4 clicks (the horizontal gray band), which is one standard deviation below the mean for passing students. The vertical gray band shows the 75-day mark.

Based on the above observations, we decided to create new virtual learning interaction variables in the form of mean total daily clicks, standard deviation of total daily clicks, total overall clicks, and the zero-day ratio; all metrics are relative to a given timeframe such as 75 days from course start. The zero-day ratio is the ratio of the number of days with zero clicks to the total number of days in the timeframe. The mean total daily clicks and total overall clicks measure how active the student is over the timeframe, and the standard deviation measures how erratic their online behavior is. We can interpret the zero-day ratio as an indication of how inactive the student has been over the period.

The number of days for the timeframe was chosen to be 75 days from course start since the first macroscopic hump in the 7-day averages plot above comes to an end around then; by the end of the first hump, we should be able to distill the difference in interactivity between the several types of students.

These interaction metrics are combined with the demographic variables, eventually leading to the comprehensive predictive model found in study 4. The following accuracy improvements from our initial base models are displayed:

	LR	RF	KNN
Student Demographics	62.1%	62.1%	63.2%
Student Demographics + VLE Features	70.6%	75.7%	78.1%

The inclusion of VLE interaction features significantly improved the accuracy of the models, indicating that a student's level of engagement is a relevant predictor of their success or failure. Moving forward, we plan to incorporate assessment features to further boost the models' effectiveness.

Study 3 (Assessment Data): How well students do in course assessments directly impacts whether they pass or fail a course. While early poor performance may be corrected with future improvement, it may also indicate negative momentum that leads to eventual failure or withdrawal. Historical assessment data are a valuable and effective feature to predict current and past students' projected outcome. Since our goal is to proactively identify at-risk students, using assessments from the beginning of the course is optimal to give the institution time to intervene and to give the student time to remediate their grade.

The assessment data includes a subset of 25,771 students and their grades earned in various courses. Students may also take the same course multiple times or take different courses as a student at Open University. Because each observation in the table is one assessment, we had to first process the data into a time series for each student. After preliminary investigation, students' mean scores strongly correspond to their `final_result` (passing with distinction, passing, failing, and withdrawing.) Scores are highest among the 'Distinction' category, followed by 'Pass,' then 'Fail' and 'Withdrawn.' The average assessment score by final outcome are as follows: Passing with Distinction 86%, Passing: 73%, Fail: 36%, and Withdraw: 23%. Students with a final score of less than 40% fail the course (Kuzilek, J., Hlosta, M., & Zdrahal, Z. 2017). Withdrawn students have the lowest scores, potentially caused by imputation of zeros for non-administered assessments.

The seven course modules (AAA, BBB, CCC, etc.), vary in assessment structure with course EEE having only four assessments and FFF having the most with 12 assessments. Additionally, the final exam scores are not included in the data. This omission does not impact the overall investigation because we are concerned with initial performance not summative performance. Other assessments are given throughout the course, either TA-scored or computer-scored.

A control chart is employed to track students' grades over time. A Cumulative Sum Control Chart (CUSUM) analysis tracks large deviations from the mean across time. A "typical" or acceptable range of values is established, and falling grades outside of the acceptable range will be aggregated in the CUSUM statistic for each time. As it pertains to this study, the distinction and passing students serve as the "typical" range of scores, using their combined mean and standard

deviation. The rationale is that students who pass have attained sufficiently high scores to be above 40% at the term's end. An assessment that falls below this "typical" range signals performance significantly below their peers. Each low score aggregates over time into the CUSUM statistic and multiple low scores result in a high CUSUM statistic and potential failure.

The CUSUM statistic C_{it} monitors dips in student's score compared to the average successful student. If a student's score is much lower than the typical score earned by the successful student, then some of the change is recorded as C_{it} for student *i* at time *t*. The standard deviation k_t for an assessment at time *t* creates a range of acceptable scores rather than a strict cutoff at mean μ_t.

$$C_{it} = \max\{0, -(x_{it} - \mu_t) + C_{it-1} - k_t\}$$

Almost all modules have identical assessment structures across their various sections. We therefore combine all sections under one course module and apply the CUSUM approach by module. An exception to this is that module BBB and DDD have anomalous sections with an incompatible number of assessments compared to other sections. While most BBB sections have eleven assessments, one section has only five. A similar situation occurs for one section of DDD. The total excluded number of students from these two sections is 2,786 students, leaving 22,985 total. Combining sections creates larger sample sizes for each module and increases the detection power of the CUSUM analysis (Wu et al. 2011).

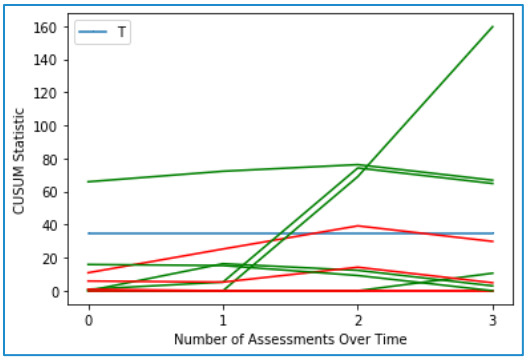
To standardize the data for the time series, zeroes were imputed for non-existent assessments so that all students had scores across the full semester. Next, every assessment per course was assigned a "typical" score, which was the mean score of passing and distinction students. The standard deviation of these students' scores signifies the normal range of scores for passing students. Lower scores that deviate from these typical zones would be cause for concern if they are significantly deviating and repetitive.

The described method of calculating CUSUM statistics was used to create a CUSUM statistic using 100% of the assessment data (all assessments taken during the semester), the first 50% of assessments, and the first 33% of assessments. While using all the assessments may yield more accurate results, our goal is to preemptively identify and support would-be failing students. Therefore, we care most about the CUSUM statistic generated by assessments 33% of the way through a course. This allows interventions to take place during the semester than may lower DFWI rates. Using a threshold (T), each student can be classified as passing or failing.

The CUSUM statistic itself does not represent any quantity related to student score; rather, a larger CUSUM statistic shows significantly poor performance while a small CUSUM statistic or zero represents results close to passable performance (generally above or near 40%).

The CUSUM statistic can be used in two ways: 1) as a classification metric on its own; or 2) as a predictor in a larger model. We do both. The results for this classification model are given in appendix Table 1. The graph to the right shows ten students CUSUM scores across the first three assessments for course GGG. When students' scores surpass the threshold (show as a blue line), students are classified as in danger of failing. The green lines indicate correctly classified students, and the red lines show misclassified students. Notice the model correctly classifies the three

green students above the threshold as ”failing,” but also does not detect three students whose lines are red but do not go over the threshold. All students who were passing are correctly classified as passing.



High accuracy is not the sole metric to be used. Sensitivity is the preferred metric because the priority is classifying struggling students even at the cost of misidentifying passing students as also failing. A threshold of 35 for the CUSUM statistic balances these metrics with a sensitivity score for predicting failing students at 85% and a precision of 75%. In practice, the model predicts 85% of failing students correctly. With higher precision at 75% we can have more confidence that the model's detection serves those who need it. A lower precision implies that the resources used to support failing students may be misappropriated.

While the CUSUM statistic alone classifies quite well, we combine the variable `Cu33` with demographic and VLE interaction variables as predictors for our comprehensive model. The accuracy improvements over our baseline (using only demographic data) are listed below.

	LR	**RF**	**KNN**
Student Demographics	62.1%	62.1%	63.2%
Student Demographics + VLE	70.6%	75.7%	78.1%
Student Demographics + VLE + Cu33	85.3%	91.2%	88.3%

As you can see, adding an assessment feature greatly improved the performance of the models, suggesting that using a combination of demographics, student activity, and early performance can be strong indicators of success or failure.

Study 4 (Comprehensive Logistic Regression Model): Early results show random forest and K-Nearest Neighbors to outperform logistic regression when comparing accuracies. Table 2 (Appendix) additionally shows sensitivity and precision as well. On all metrics, random forest performs best.

We care most about sensitivity, that is, how well the model classifies failing student out of the total actual failing cases. This is balanced with precision, the percentage of true failing cases out of the total classified failed students. Though medical tests have high sensitivities above 95% to detect diseases, our classification does not have life and death consequences. Therefore, we aim for a sensitivity between 80% and 90%. Precision is also important because higher estimates improve the chance that interventions go to those who need them. We have determined that optimal precision should not drop below 50%. Sensitivity measures the proportion of positive cases (e.g., students at risk of underperforming) correctly identified by the model. The primary goal of such a model could be to accurately identify as many students as are genuinely at risk of underperforming as possible. This is crucial for providing timely intervention, support, and resources to help these students improve their performance. In short, high sensitivity ensures that fewer at-risk students are missed by the model.

Even though the random forest and KNN models perform best, they may be prone to overfitting. Moreover, the logistic regression could be improved by further tuning the threshold value (by default the threshold is 0.5). In terms of complexity, logistic regression may be preferred for easier interpretability and low complexity.

Final Model Results (Including Thresholds): Using only demographic variables and a threshold value of 65%, classification accuracy was 56.3% with a sensitivity of 76.9% and precision of 47.8%, which is not much better than guessing. Therefore, modeling solely based on these features is insufficient to provide timely targeted intervention.

In addition to demographic information about students, Open University provides data on the frequency of student virtual learning environment interaction. Earlier we created new variables: `mean_clicks`, `stdv`, `total_clicks`, and `zero_day_ratio`, which summarize total student engagement. Because `mean_clicks` and `total_clicks` are naturally linearly dependent, we exclude one of them to avoid multicollinearity. An analysis of VIF corroborated this high correlation. Adding `stdv`, `total_clicks`, and `zero_day_ratio` into the model results in a classification (with a threshold value of 70) slightly better with 59.9% accuracy, 85.4% sensitivity and 50.5% precision.

Next, we add assessment data in the form of a cumulative statistic (Cu33) as described previously. Recall that the cumulative statistic tracks a student's progress in a course with small values indicating acceptable passing performance, while large values represent frequent poor performance. The variable is `Cu33`, which represents the performance of each student on the first 33% of the assessments. Adding this variable improves model performance. Notably the precision jumps from 50% to 78% and accuracy increases from 60% to 84% while maintaining a sensitivity of 85%. Increased accuracy indicates high ability to classify students overall, while high sensitivity and precision indicate most failing students are identified. The results are summarized below:

Feature Selection	Threshold (T)	Accuracy	Sensitivity	Specificity	Precision
Demographic Only	65	0.5631	0.7690	0.4213	0.4777
Demographic + VLE Interaction	70	0.5987	0.8532	0.4235	0.5046
Demographic + VLE Interaction + Assessment	70	0.8425	0.8535	0.8350	0.7807

Lastly, we compare logistic regression results across different thresholds for classification in Table 3 (Appendix). As we increase the threshold for classification for passing, accuracy remains relatively stable at about 84% - 85%. However, the sensitivity for detecting fails/withdraws increases and precision decreases. A higher threshold of 0.70 compared to the typically used 0.50 increases sensitivity from 75% to 85% and decreases precision from 90% to 78%. This analysis of tuning the logistic parameters yields better sensitivity, but does not beat KNN or random forest when comparing classification metrics.

Conclusion: We have shown that the DFWI rate is significantly influenced by four primary factors: student demographics, academic readiness, course engagement metrics, and assessment performance. We have also shown that each factor independently affects the failure and withdrawal rates. Students that come from poorer economic backgrounds tend to perform worse than those that come from wealthier backgrounds. Students that are more active in the virtual learning environment tend to have better outcomes than those that are less active, and students that perform well on the earlier assessments tend to have better outcomes than those that perform poorly on the earlier assessments. These insights have led us to some actionable interventions:

Intervention 1: To support students in more deprived areas, as indicated by IMD less than 50%, weekend tutoring and study groups can be offered in deprived area local libraries.

Intervention 2: After the first 75 days, a student that falls below an average of 4 clicks per day should be offered to the instructor/TA and queued for personal contact.

Finally, we have shown that incorporating assessment and interactivity data, in addition to student demographics data, in our models improves predictive performance of our models. Our comprehensive prediction model enables faculty and staff to predict student success in a timely manner for actionable intervention, which leads us to our final and ultimate actionable intervention:

Intervention 3: **A comprehensive model**. Once a certain amount of time has passed from course start—for instance, 75 days—and students have taken the first 33% of the module assessments, all student data, including demographic, interaction, and assessment data, are run through our comprehensive predictive model, and the instructor is automatically presented with a list of at-risk students to reach out and provide support to.

Further Investigation: We saw that running prediction models using demographic data, assessment data, and interaction data together led to superior performance over running the models using each set of data separately; predictive power is emergent, so to speak. Understanding this phenomenon is worthy of further exploration. We also tested other models such as random forest and KNN and saw promising signs of improvement in performance relative to logistic regression. Further fine-tuning of models to attain optimal levels of accuracy, precision, and recall is worth exploring. Finally, alternative and more advanced features could have been engineered from the assessment and interaction data, leading to better performance of our comprehensive model.

References

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education, 17(3). https://doi.org/10.1186/s41239-020-0177-7

Custer, S., King, E. M., Atinc, T. M., Read, L., & Sethi, T. (2018). Toward Data-Driven Education Systems: Insights into Using Information to Measure Results and Manage Change. Center for Universal Education at The Brookings Institution. Retrieved from ERIC database. (ED583026). https://eric.ed.gov/?id=ED583026

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

Smith S.E., Tallentire V.R., Spiller J., Wood S.M., Cameron H.S. "The educational value of using cumulative sum charts," Anaesthesia. 2012 Jul;67(7):734-40. doi: 10.1111/j.1365-2044.2012.07100.x. Epub 2012 Mar 15. PMID: 22420772. The educational value of using cumulative sum charts - PubMed (nih.gov)

Wu, Z., Yang, M., Khoo M. B.C., Castagliola P., What are the best sample sizes for the Xbar and CUSUM charts?, International Journal of Production Economics, Volume 131, Issue 2, 2011, Pages 650-662, ISSN 0925-5273, https://doi.org/10.1016/j.ijpe.2011.02.010

Appendix

Table 1: Best Results for CUSUM Across Different Thresholds and Course Modules

Course	Threshold (T)	Accuracy	Sensitivity	Specificity	Precision
All	35	0.8228	0.8500	0.8044	0.7459
AAA	35	0.8213	0.6571	0.8755	0.6354
BBB	35	0.8102	0.9138	0.7433	0.6971
CCC	35	0.8055	0.9286	0.6780	0.7491
DDD	35	0.7904	0.6869	0.8781	0.8268
EEE	35	0.8325	0.5331	0.9503	0.8084
FFF	35	0.8511	0.9599	0.7722	0.7535
GGG	35	0.8414	0.8411	0.8415	0.6770

Table 2: Model Comparisons

Student Demographics +VLE + Cu33	Accuracy	Sensitivity	Precision
Logistic Regression	0.853	0.748	0.874
Random Forest	0.912	0.836	0.941
K-Nearest Neighbors	0.883	0.806	0.897

Table 3: Logistic Regression Across Different Thresholds, using Demographic + VLE Interaction + Assessment

Threshold (T)	Accuracy	Sensitivity	Specificity	Precision
45	0.8504	0.7253	0.9364	0.8870
50	0.8533	0.7491	0.9249	0.8730
55	0.8543	0.7717	0.9111	0.8566
60	0.8521	0.7953	0.8912	0.8342
65	0.8494	0.8232	0.8675	0.8104
70	0.8425	0.8535	0.8350	0.7807
75	0.8225	0.8854	0.7791	0.7340
80	0.7867	0.9164	0.6805	0.8943