Classification and Features of Student Success in Online Programs

Jacob Javier

Springboard Data Science Career Track

Classification and Feature of Student Success in Online Programs

Since 2020, there has been a major educational shift out of the classroom and into virtual learning environments (VLE). In the past decade alone, the proportion of students enrolling in online education nearly quadrupled (Hamilton 2024). Although online education provides students with more flexible access to the curriculum and educational administrators, educators find it difficult to gauge how well the material is transferring to their students. The separation of student to educator removes much of the nonverbal communication from the classroom, only leaving hard data behind.

Thanks to Dr. Kuzilek and their colleagues through The Open University (2015), online academic program data is being made available to understand driving factors for successful students in VLEs. The data provided could simultaneously be used to train classification models that could help administrators predict whether students are going to meet the Student Learning Outcomes. Although the models trained out of this data set may be out of date, they may provide the framework to build more complex and robust tools for other programs or times.

Data Wrangling

The Open University (2015) data set depicts the outcomes of seven online courses (modules) from The Open University. Seven separate comma separated value (csv) files were provided as the complete data set. In total, there were 32953 students that worked on 206 different assessments require various levels of interaction with the VLE. The data set provides the demographics of the students, their individual interactions with each activity as part of the assessments, and the classification of how they ended the course: Distinction, Pass, Fail, or Withdrawn. Distinction is a higher classification than passing. No final scores were provided in any of the csvs.

The assessments csv characterized the assessments the students were being scored on. Characterizations of the assessments included which module they belonged to (‘code_module’), assessment identifiers (‘id_assessment’), the due date (‘date’), and the impact of the assessment on the overall grade (‘weight’). The only feature that was used from the assessments csv was the final submission date. There were a few assessments that were missing submission dates that were imputed with the maximum submission date. The maximum submission date represents the last day of the course since all dates are formatted as number of days since the start of the module. By imputing the last day of the course, the students are assumed to not be able to submit past the end of the module. All other features were either used for merging the DataFrames or dropped all together.

The courses csv provided the module identifiers as well as the length of the course. All features from this DataFrame were dropped after merging because the focus of this model is features directly related to or within the students’ control.

The studentAssessment csv showed how well each student scored on their respective assessments (‘score’) and when they turned in each assessment (‘date_submitted’). Because some students withdrew from the courses, not every assessment was completed by every student. Administrative features like whether the assessment was transferred, or the unique identification number (ID) were dropped prior to model construction. The DataFrame was missing about 0.09% of the observations' scores. With such a small proportion missing of a large dataset, the observations were dropped.

The studentInfo csv showed the demographic information of each student. Students were only distinguished by a unique ID (‘id_student’). Even the students’ ages were classified within age bands (‘age_band’) to further protect their identities. Administrative features like ID were dropped prior to model construction and analysis. The main features that were provided within this DataFrame included the final result (‘final_result’) of the student with the module, the highest level of education attained (‘highest_education’), and the number of additional credits being studied by the student (‘studied_credits’), among other demographic information. The Index of Multiple Depravation (IMD) band establishes the financial bracket of the students. Missing values in the IMD band feature (‘imd_band’) were imputed with ’20-30%’ which accounted for 12% of the feature’s data initially.

The studentRegistration csv provided the date each student registered for each module; and if they withdrew, what date they unregistered. Only the registration date (‘date_registration’) was retained of all the features to see if there was a relationship between when a student registered with how well they did in the course. The dates were all relative to the start of the module. Negative dates indicated that the student registered prior to the start of the course.

The studentVle csv was used for the number of days the students interacted (‘date’) with the VLE and how many clicks they performed on each of these days (‘sum_click’). All other features were administrative in nature that were only useful for merging. The number of clicks were of special interest as a quantification of student engagement with the VLE.

Finally, the vle csv was used to characterize the courses with the various VLE activities. Only the types of activities (‘activity_type’) were retained after merging to see if any part of the VLE was associated with how students finished the course.

All duplicates were dropped and the only feature that was simply renamed was ‘date_submitted’ to ‘assessment_duration’. The date provided by this feature can be used as a duration because all dates are relative to the start of the module.

Aggregations

While 'date' is a duration length relative to the start of the module, it only served as a marker for each day of interaction with each activity the students interacted with per assessment so it was aggregated as a count of the number of days interacting with the material to quantify the interaction with each activity type (‘days_active’). Similarly, 'sum_click' was aggregated to the total number of clicks per activity (‘total_clicks’).

The number of clicks were further aggregated as the total number of clicks per assessment (‘clicks’) and the average number of clicks for each assessment (‘mean_clicks'). The number of active days were also aggregated per assessment into the sum total (‘total_active’) and average (‘mean_active’). As two distinct quantifications of student interaction with the material.

The scores were also averaged (‘mean_score’) for each assessment to represent the consistency and level that students operated on. The ‘assessment_duration’ was also averaged (‘mean_assessment_length’) and maxed (‘max_assessment_length') to represent how long students had with the material.

Outliers

Outliers were identified as the most extreme 5% of data for each of the continuous features (‘studied_credits’, ‘weight’, ‘sum_click’, ‘date_registration’). Reducing the number of observations to the middle 95% of values maintains the distribution quality while limiting the effects of the outliers present. In some instances, the outlier of a distribution was hundreds of units above the next highest value. To minimize the effects on the model’s scaler, these outliers were removed from analysis.

Data Exploration

The two main goals outlined for this model is to identify the key drivers of how students ended the course and create a classification model that enables educators to predict student trajectories. In relation to these goals, four questions were used to initialize understanding student VLEs.

1. How does the average score for each assessment differ between the final results?
2. How does the material interaction for each assessment differ between the final results?
3. Is there a difference between activity types that determine the final results?
4. Does education interact with students' final results?

How Does the Average Score for Each Assessment Differ Between the Final Results?

One of the most important predictors that would dictate a student's final result is how well they scored on average with their assessments. A student that scores higher on average per assessment should place higher with a pass or even a distinction final result since they would have met the minimum passing grade per assessment at least.

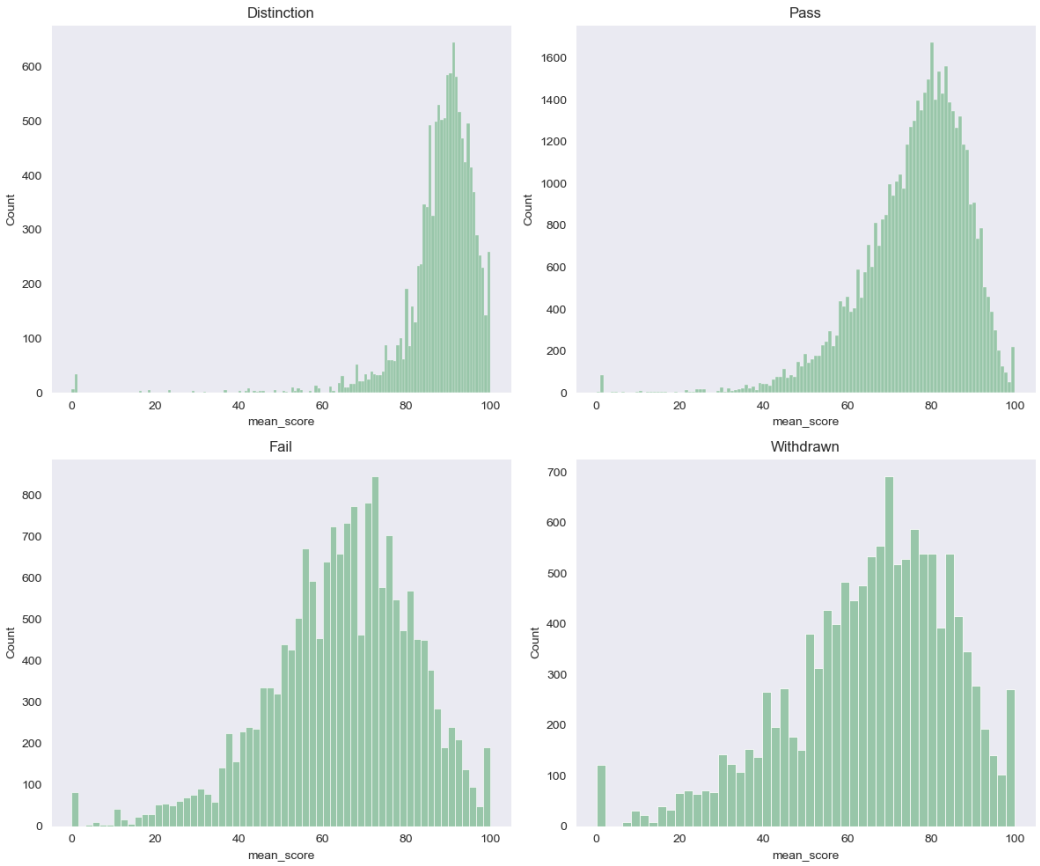

Fig 1. Distribution of the average score for each assessment submitted by all students.

Lower ranked student results are more normally distributed than distinguished or passing students (Fig 1). Despite having a difference in a couple of points there is a significant difference (f-stat = 8332.65, p-value = 0.0) between the distribution means of students who failed (mean score = 65.14) and withdrew (mean score = 63.74). Even with Bonferroni correction, all four classifications of students were tested to not be like one another (f-stat = 8332.65, p-value = 0.0).

From the initial testing, the average score for each assessment appeared to be a driving factor in classification. Students who, on average, scored higher were more likely to pass or be distinguished than those who scored lower.

How Does the Material Interaction for Each Assessment Differ Between the Final Results?

Increasing the amount spent with the material may indicate a student that is taking special care with the assessment or a student that is dedicated enough to see it through. Longer amounts of time could also indicate a concept that is a little too difficult to grasp or some confounding factors outside of a student's control. Regardless, the longer the student took to turn in the assessment, the more opportunities to interact with the material, the student had.

To better understand this question, the material interaction was broken down into the number of interactions ('mean_active') and how long each assessment took on average ('mean_assessment_length'). The number of interactions is a more robust indicator to of how proactive a student was on their assessments. The average length of time is a passive indicator that encompasses students who are struggling with the material, taking extra time to perfect their assessment, taking breaks, or even procrastinating. The two features were assessed separately because very little of the variances between the two features were explained by the other (r = 0.0785).

There was no significant difference (f-stat = 2.25, p-value = 0.0796) found between any of the four group results in terms of interactive frequency ('mean_active'). The average time spent on assessments ('mean_assessment_length') between the four groups did significantly differ between one another (f-stat = 6453.60, p-value = 0.0).

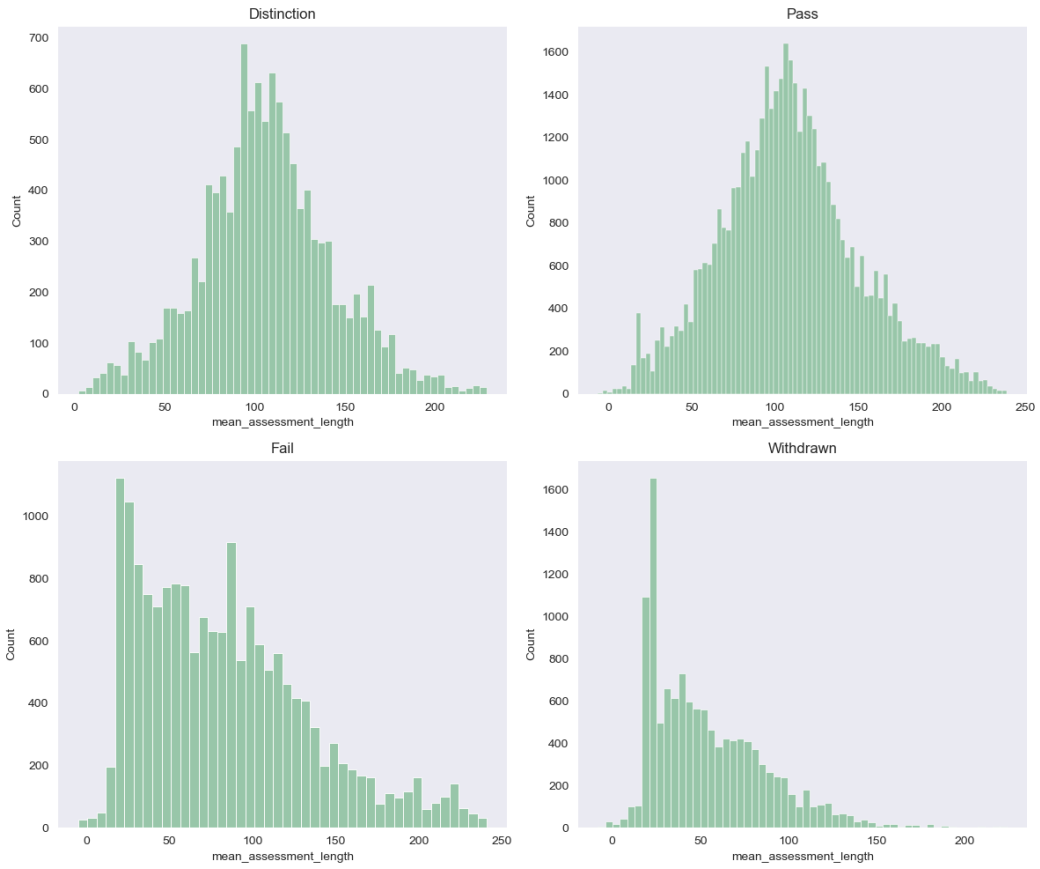

Fig 2. Distributions of the average assessment duration for each student.

Students who at least passed (mean duration = 108.87 days) took at least 20 more days for each assessment than those who failed (mean duration = 83.11 days, Fig 2). Distinguished students (mean duration = 105.99 days) took less (f-stat = -6.92, p-value = 0.0) time to complete their assessments than students who passed, which may be a sign of grasping the material more efficiently. Students who withdrew (mean duration = 53.93 days) may have had

significantly less time due to only the early assessments being turned in or their end dates limiting their assessment lengths.

Is There a Difference Between Activity Types that Determine the Final Results?

By exploring the activity relevance in relation to student success or failure, resources can be added to support or remove an inefficient activity type. Among the seven assessments, there were 19 different activities in the VLE that students used to interact with the material. Over a fifth of the interactions were with the home page. The other activities that comprised the top 76% of interactions were subpages, Open University content, forums, and external resources. The course seemed to have a from the top activities, it appears that much of the course was focused on non-verbal discussions and transactions of the material. This would make reading student's level of engagement difficult for educators in the program.

Fig 3. Density plots of how often all activities were interacted with for each final result.

Almost all activity types had passing students as the greatest value (Fig 3), which may be due to this classification representing over half of all the students. The only activities that students who withdrew did not interact with were folders and the shared subpage (Fig 3), which may be because those activities became available later in the course and would not be accessed unless the module was completed. Another interesting note is that only students who failed or

passed accessed the shared subpage (Fig 3). The shared subpage may have been an optional activity that distinguished students chose to ignore in favor of point efficiency.

The activity types did express some dependence with the final results (p-value = 0.0) so the model is expected to use this as a predictive feature. Interaction with the different activities would make sense to dictate placement at the end of the module since too much interaction with lower weighted activities would be an inefficient use of the time.

Does Education Interact with Students' Final Results?

Higher degrees of education are expected to have had more exposure and training within an academic setting. If there is a distinction between students at varying levels of education, this would enable administrators to develop support systems for those students in future cohorts.

Table 1. Conversion of the United Kingdom educational values in the highest education feature to United States of America equivalents.

UK ('highest_education')	USA
No Formal quals	Did no complete high school
Lower Than A Level	Completed high school and partial completion of an Associates degree
A Level or Equivalent	Advanced Placement high school diploma or the first year of a Bachelor's Degree
HE Qualification	Bachelor's Degree
Post Graduate Qualification	Postgraduate Certificate or higher education

Because this model is being constructed in the United States of America (USA), USA educational equivalents have been translated from the UK educational system (Table 1). There was some dependency (p-value = 0.0) between the highest education level and the final student placement. Similar to activity types, passing students represented almost half of the plot density for each education level (Fig 4). As the level of education increased, so did the quantity of distinguished; and the number of failed students lowered (Fig 4). The proportion of withdrawn students varied very little between educational levels (Fig 4).

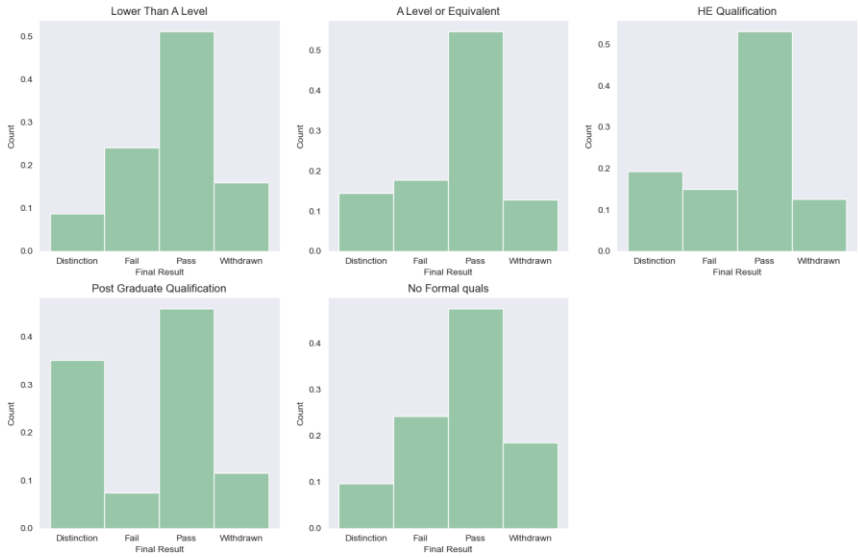

Fig 4. Density plots of highest education levels for each student placement.

The increased educational accreditation may be associated with higher placement due to more time spent with academic resources. Students who have immersed themselves in academia longer may have more training in learning techniques to learn and complete the material more efficiently.

Modeling

Four models were constructed to determine which would provide the best classification for the data set. The simple decision tree model (dt) was the first model to train as a baseline comparison to all other models. The dt model is a supervised machine learning (ml) algorithm that can be used as a classifier that selects the tree based on the lowest mean squared error. For this model the tuning hyperparameters were the splitting method ('criterion'), how many layers the trees could have ('max_depth'), and what the minimum number of samples in each leaf allowed ('min_samples_split'). The latter two hyperparameters were included to prevent overfitting.

Next the random forest (rf) model was constructed because it can combine multiple decision trees for a more robust model. Both the dt and rf models were of particular interest because they were able to identify the driving factors of the final results classification. Like the dt model, the number of features present in each tree ('max_features') was determined to be best set to the square-root of the number of features. The only additional hyperparameter that was tuned was the number of trees that were made ('n_estimators').

K-nearest neighbors (knn) and logistic regression (lr) models were also trained to more probabilistically classify the final results. Neither of these models provides feature importance identification, but use different methods than the decision trees for classification. The knn model was hypertuned to the number of neighbors (‘n_neighbors’), the weight distribution

('weights'), and the algorithm to compute the nearest neighbors ('algorithm'). The lr model was hypertuned to the penalty calculated to each feature ('penalty'), the type of algorithm used for optimization ('solver'), and the regularization strength ('C').

Encoding

To prepare the data for modeling, the data was encoded since the models cannot process strings efficiently. Student gender ('gender') and disability status ('disability') were binarily encoded to '0' and '1'. Data from this online program did not include any other gender identities so a binary classifier was most appropriate for this feature. Similarly, disability status only checked for whether a student registered as disabled, not the nature of the disability.

The features that were ordinally encoded were education ('highest_education'), IMD band ('imd_band'), age ('age_band'), and final result ('final_result'). The IMD and age bands expressed a clear order of progression. For education, the order is the amount of time spent in academia. The final results were also encoded as a progression of rank with 'Distinction' as the top, and 'Withdrawn' in the bottom. Distinction was deemed the highest rank between the four because it is a special honorific beyond passing. Withdrawn was considered the bottom because students who withdrew did not make it to the same standard as the other students to determine their ranking.

The region students were from ('region') and assessment activities ('activity_type') were nominally encoded using value counts. To ensure that these features were not underweighted during analysis, they were encoded by their feature value counts. Although the encoded values ranged to the thousands, proportionally, their weights were still representative of the distribution.

Scaling and Splitting

Because the data was ordered by assessment originally, the data was shuffled prior to splitting. Additionally, the data was stratified to ensure a proportional representation of the classifications were present in both the training and test sets. Students who passed severely outweighed any other classification. A quarter of the data was reserved as the test set.

The data was scaled using a StandardScaler to ensure all distributions have similar variances for the parametric-based models. The StandardScaler was chosen above a Normalizer or RobustScaler because the outliers had been dealt with during data wrangling.

Model Results

Model classifications were compared based on their F1-scores to ensure the most consistent identification of true positives. The F1-score was also chosen because it was accessible by all four classifiers and is more robust to the class imbalance than accuracy.

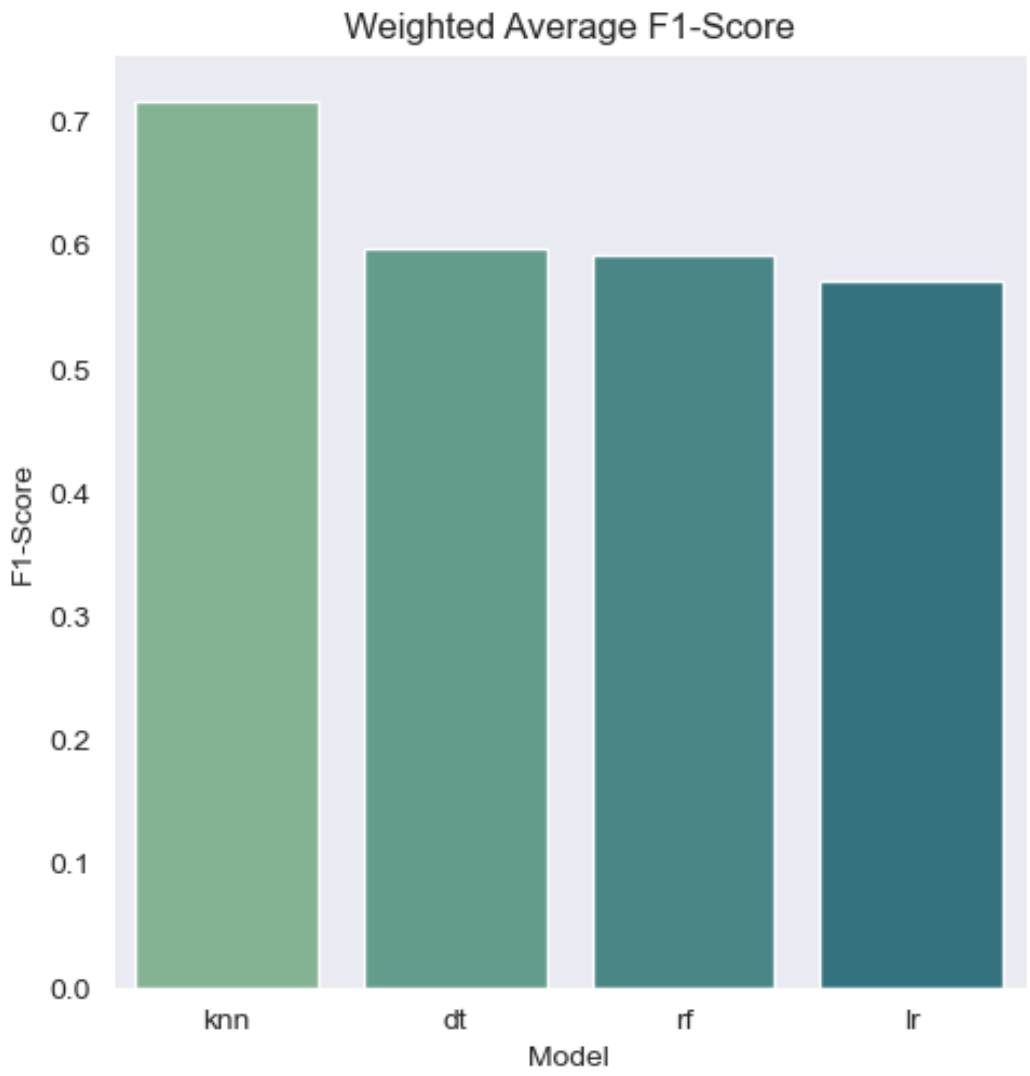

Fig 5. Weighted F1-scores of the knn, dt, rf, and lr models.

The best model for classifcation was the K-Nearest Neighbors model (Accuracy = 0.72, Weighted F1-Score = 0.72, Fig 5). Passing students were the easiest to classify for all four models; to the point of mistaking most of the distinguished students for passing classifications. The class imbalance may have severely impacted the prediction of the models to almost default to the passing classification. The knn model is the best model to classify students, however with an accuracy and weighted F1-score of 0.72, there are more tunings required before deploying it as an educational aid for online educational administrators. Being able to predict a student's trajectory within 72% is very risky. For future iterations of this classification tool, under sampling the passing students to accommodate the class imbalance may improve the model metrics across all models.

Although the knn model provided the best classification, the best decision tree classifier is needed to determine feature importance. Both the dt (Accuracy = 0.62, Weighted F1-Score = 0.60) and rf (Accuracy = 0.64, Weighted F1-Score = 0.59) were very similar when comparing

their weighted f1-scores (Fig 5). The rf model was chosen for feature importance over the dt model because it scored higher in accuracy than the dt model.

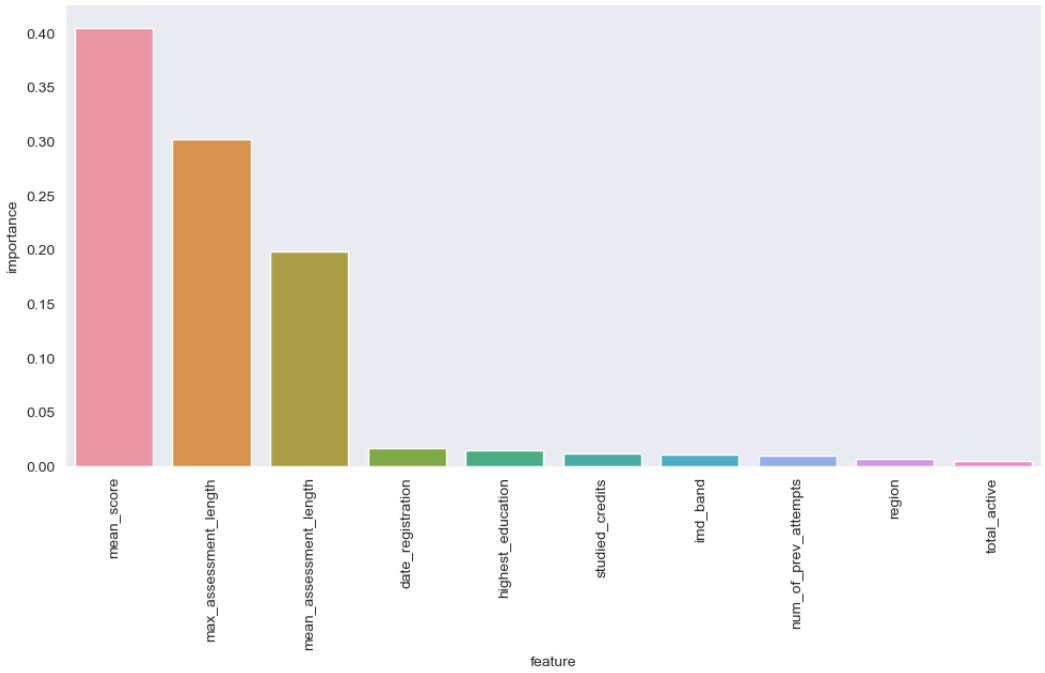

Fig 6. Random forest feature importance predicting the final results of the courses.

As anticipated, the average score for each assessment was the strongest predictor ('mean_score' Importance = 0.40) of how a student would end the course (Fig 6). Additionally, the amount of time a student spent on their assignments ('max_assessment_length' Importance = 0.30, 'mean_assessment_length' Importance = 0.20, Fig 6) were also strong indicators of how they would be classified by the end of the course. Between these three features, 90% of the classification was predicted. The major distinction between the top three features and the rest may indicate that the program provides equitable enough resources to its students that only their work ethic determines their success (Fig 6).

Conclusion

The goal of these models were to create a classification tool that educators could use to predict student trajectories without easy access to them; and to identify the driving factors of how students succeeded or failed. The main determinants identified for this program were how well students scored on average in their assessments and how long they spent completing those assessments. The program indicated through feature importance that it provides equitable resources that enables all demographics to pass or fail by their own work. All demographic features explained less than 10% of the how students would end the course which would be a great indicator of an equitable program.

This project was able to correctly classify, at best, about 72% of the students. The easiest students to classify were passing students who comprised over 50% of the data. To create a more robust model, the class imbalance between how students ended the course needs to be reduced through resampling techniques. The imbalance can be compensated in future studies by subsampling passing student observations or aggregating data across multiple years.

From these findings, administrators of this course may need to invest into maintaining the balance of equitable opportunities for all students and providing sufficient academic support to ensure high scoring. In future models, understanding the features that dictate the higher scores would be crucial to identifying additional resources needed for development. Additionally, more studies are needed to contextualize how assessment duration is associated with students passing or failing the course. In terms of the students who completed the course, they would have had access to more of the scoring credits than students who withdrew so it could be an indication of students simply participating. Other possible explanations are measuring the level of care or dedication to a particular assessment or students not being able to understand the assessments clearly enough. During exploratory data analysis there was an association between the highest level of education and how students finished the courses. Reanalyzing the data controlling for education rather than student or assessment will be a good indicator of how education scales with course placement. Although determining how activity types interacted with the final results became a secondary objective through exploratory data analysis, it was one of the least important features in prediction. Running additional PCA tests would help reweigh the types of activities to better determine if specific activities were more deterministic for students than others.

From this model, driving factors of student success were able to be identified, but the accuracy and consistency of the models would prevent them from being deployed as classification tools. After tuning and resampling, these model results have the potential for equipping online educators to better engage their students.

References

Hamilton, I. (2024, May 31). 2024 Online learning Statistics. Forbes Advisor. https://www.forbes.com/advisor/education/online-colleges/online-learning-stats/

Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z. and Wolff, A. OU Analyse: Analysing At-Risk Students at The Open University. Learning Analytics Review, no. LAK15-1, March 2015, ISSN: 2057-7494.
source: http://archive.ics.uci.edu/ml/datasets/Open+University+Learning+Analytics+dataset