Early Intervention Modelling for Withdrawing Students at Different Course Intervals

DGLL43

ABSTRACT

Online learning has been rapidly becoming more popular within higher education in recent years, providing advantages over in-person learning in regards to flexibility and cost. Where online learning platforms fall short is in learner retention; large proportions of learners fail to complete or even withdraw from online courses. In this work we examine the task of early prediction of learner dropout at different points in a course's lifetime using a subset of OULAD. We evaluate a set of temporal based indicators which map to higher level learner qualities. These demonstrate a greater correlation with learner dropout than demographic features and provide improved intepretability for non-technical stakeholders. Using Random Forest classification we produce a fine-tune model that achieves a mean F1 score of 0.766 for dropout prediction with only 25% of the temporal data, increasing to 0.815 at 100%. Finally we provide a comparison between ensemble based methods and individual classifiers for their effectiveness in early prediction for learner withdrawal

CCS CONCEPTS

• **Applied computing → Education**; • **Computing methodologies → Supervised learning**.

KEYWORDS

Early intervention, machine learning, at-risk learners, OULAD, prediction, dropout

1 INTRODUCTION

Within recent years the use of online learning through platforms such as Virtual Learning Environments (VLE) has become increasingly more common in higher education and as a result of COVID-19 has even become the dominant form of teaching in the short term. VLEs provide a highly scalable and affordable approach to learners that is far more flexible than in-person teaching. As life slowly returns to normal it is unlikely that online teaching will return to pre-COVID levels of uptake and will more likely continue to be an important avenue of learning for higher education institutions.

COVID-19 enforced learning notwithstanding, research suggests that VLE retention rates fall drastically short of those for in-person teaching. The work of Jiang & Kotzias [7] suggests that as little as 7% of learners complete courses presented through VLEs. This provides substantial motivation for addressing the task of dropout prediction.

The advantage of online learning from an analytical viewpoint is the capture of student information and their interactions with the VLE over the duration of the course. This data in turn can be coupled with machine learning to develop a predictive model for learners at risk of withdrawing from the course. While such a model provides value to course designers by predicting the results of an alteration to the course, the most value is derived during the learning period by enabling teachers to intervene before a learner leaves the course. Teachers are able to provide personalised support to those learners most likely to require it.

Early-intervention predictions present a challenging task, requiring models to generate predictions based on partially complete courses since identifying that a learner will withdraw at the end of the course will be too late to provide meaningful intervention. Taking place within the domain of Learning Analytics, the stakeholders of such a system represent the intersection of several different fields. Those with a background in Education may lack the technical experience to leverage knowledge from features used for prediction that do not translate naturally to higher level semantic concepts. Therefore it is vital to consider the interpretability and explanability of such a system in tandem with prediction performance in order for stakeholders such as teachers to provide meaningful support.

With this in mind this work makes three major contributions. The first is the engineering and evaluation of assessment and behavioural features as representations of higher level concepts,, such as learner motivation, as indicators of learner dropout. Second is the investigation of a set of machine learning algorithms and subsequent generation of a predictive model using these features that identifies learners at risk of dropout at different intervals in course duration. Finally we present an experimentally tuned ensemble predicting for early-intervention predictions.

2 RELATED RESEARCH

Prior work has already demonstrated the potential of machine learning (ML) for developing a model for early prediction of at-risk students. [9] utilises three ML algorithms to produce a prediction model for dropout; feed-forward neural networks, support vector machine (SVM), and probabilistic ensemble simplified fuzzy ARTMAP. Incorporating data concerning learner demographics, assessments and interactions their method a 75–85% accuracy in overall student classifications for two out of the three learner courses it is applied to. [6] applies Decision Trees, Naive Bayes, SVM and Logistic Regression to a substantial dataset of 9900 students. Logistic Regression achieves the highest accuracy for students who eventually withdrew from the course at 94.2%.

The work of Costa *et al.* [4] applies four ML algorithms to a small dataset of several hundred students. Decision Trees achieve an F-measure value of 0.82 with respect to dropout prediction when applied to the data collected in the first week of a 10 week course. [4] also highlights the importance of feature engineering, using the Information Gain algorithm [10] to select the most distinguishing attributes followed by the Synthetic Minority Oversampling

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

Technique (SMOTE) algorithm [3] for class-balancing via data augmentation.

[2] applies logistic regression to grades achieved throughout several e-learning courses. The models achieve above 90% accuracy for each course when at week 10 of 20. When implemented in conjunction with a tutoring plan based on the identification of learners at risk of dropping out, the dropout rate was reduced by 14%. While impressive there is limited evidence of generalisability with models trained upon only 100 students across all of the courses. [1] evaluates the quality of at-risk learner predictive models produced when trained upon OULAD [8] learner data at {20%, 40%, 60%, 80%,100%} course duration by date. Random Forest achieves the best results with F-measures of {0.59,0.79, 0.84, 0.88, 0.90, 0.91}. The task first merged *Distinction* with *Pass* results and *Fail* with *Withdrawn* results, therefore addressing a slightly different task to dropout prediction.

[5] transforms raw data from OULAD [8] into behavioural indicators in order to provide a semantic aspect and improve the interpretation of such features for non-technical stakeholders. Applying Random Forest achieves an F-measure of 0.85 on a subset of OULAD [8] and while this is using the entire course duration it does illustrate the potential of incorporating higher level features while maintaining model performance.

3 METHODOLOGY

3.1 OULAD

We use the Open University Learning Analytics Dataset (OULAD) [8] produced by the Open University. The dataset contains data collected for years 2013-14 with 32,593 students, across 22 module-presentations of 7 courses. Most importantly for this work the dataset contains two sets of temporal data; the students' daily interactions with materials in the VLE, and assessment scores with the date they were submitted. [12] argues that clickstream data in particular is more accurate and objective than alternate, self-reported data for capturing learner behaviour. This data alongside the non-temporal student demographic information enables the evaluation of our selected methods in prediction learner dropout. We focus on the *2014B* presentation of the *BBB* course for our exploratory analysis.

3.2 Feature engineering

We convert ordinal categorical data into encoded sequence labels and encode the remaining categorical data with standard label encoding. We evaluate the quality of models produced using four different proportions of the course data divided by time {25%, 50%, 75%, 100%}, in tandem with the non-temporal data. Expanding upon the work of [1] we generate aggregate features to capture learner performance of the period in a lower-dimensional space with greater interpretability for stakeholders. Relative Score (*relative_score*) is the weighted sum of assessments submitted during that period. Raw Score (*raw_score*) is the unweighted sum of assessment scores. Late assessment score (*late_score*) describes the number of assessments submitted late.

We incorporate behavioural indicators first demonstrated in [5] to dually improve the accuracy of the model produced and to increase the knowledge that can be leveraged about learner behaviour by stakeholders. We use the proportion of assessments submitted at that point as a measure of perseverance (*assess_complete*). [5] uses the clickstream behavioural patterns to generate three measures of learner commitment or engagement; collaborative (*collab_click*), course structure (*course_struct_click*), and course content commitment (*course_content_click*). Each is a summation of clicks with a set of activities. Collaborative commitment describes the willingness of a learner to share knowledge and engage with others, interacting in forums or wikis. [11] found that low interaction with the course home page is an indicator of potential dropout and [5] builds upon this with course structure commitment as a sum of interactions with pages such as the home page and glossary. Course content commitment is the sum of interaction with course materials. Evaluation commitment (*eval_click*) describes a learner's engagement with assessments. An overall sum of all clicks (*sum_clicks*) is used as a indicator for a learners motivation.

3.3 Class imbalance

The final result is encoded as binary labels, {*Distinction, Pass, Fail*} are encoded as 0 and {*Withdrawn*} is encoded as 1. Understandably the dataset is imbalanced with dropouts as the minority class. Machine learning methods applied to imbalanced datasets can result in a bias towards the majority class as a quicker way to improve the overall prediction accuracy rather than learning to accurately model the data. We employ SMOTE in order generate synthetic samples for the minority class to balanced the dataset. SMOTE interpolates between existing vectors as a more intelligent method for augmentation in comparison to simple oversampling or undersampling.

3.4 Machine Learning

We employ five different ML algorithms for modelling learner dropout over different proportions of course completion; Random Forest, ExtraTrees, Logistic Regression, SVM, and Gradient Boosting due to their demonstrated ability for modelling this task in prior work. Random Forest fits a set of tree classifiers to sampled subsets of the training dataset, taking the average of their individual predictions to generate a classification. ExtraTrees is a similar approach to Random Forest but where Random Forest subsamples the dataset with replacements, ExtraTrees uses the entire dataset to fit the individual trees. Logstic Regression is a linear classifier that aims to find the best weights for the logistic regression function often by maximising the log-likelihood function for all training samples. The SVM aims to locate the hyper-plane in the feature space that separates the two classes in order to classify them. Gradient Boosting again uses trees, trees are fitted to the training set and greedily selected based upon performance, gradient descent is then used to identify which trees to add to the model that minimise the chosen loss function.

3.5 Evaluation Criteria

We utilise a set of metrics built upon the true and false positives and negatives for evaluating model performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy describes what proportion of the predictions were correct, the sum of True Positives (TP) and True Negatives (TN) divided by the number of predictions.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Precision and recall are particularly appropriate for tasks such as these with imbalanced datasets. Precision describes the proportion of predicted positives that are TPs. Recall describes what proportion of ground truth positives have been predicted. A model that has predicted only the majority class could be high in accuracy but this bias would be highlighted in low precision.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

F1 Score is the harmonic mean of recall and precision. In order to highlight disparities between the classes we the mean of the class-wise F1 scores as an effective metric for the performance of a model with our imbalanced dataset.

To generate training and testing splits we use K-fold cross-validation with a k value of 5. This divides the dataset into k folds, k-1 folds are used for the training set and the remaining fold is used for testing. This enables the most accurate model performance evaluation to be undertaken using the entirety of the dataset for training and testing a different points.

4 RESULTS

4.1 Feature Correlation

We initially evaluate the relationship between our extracted features and the occurrence of dropout. To provide a meaningful evaluation for non-ordinal categorical data we calculate the Uncertainty Coefficient for those features (Fig. 1). Only *disability* has even a minor reduction in uncertainty for dropout. Learners with a disability are more likely to dropout of the course whereas able-bodied learners are more likely to complete this course. This suggests that the course may not always be able to adapt effectively to the needs of disabled learners or alternatively that these learners are more likely to encounter situations that result in them withdrawing from the course.

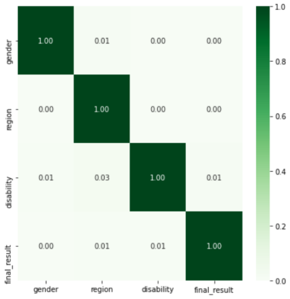

Figure 1: Uncertainty Coefficients for non-ordinal categorical learner demographic features

Calculating the Pearson correlation for the remaining demographic features (Fig. 2) we find that *studied_credits* has the greatest correlation at 0.16. Learners who have studied or are studying a higher number of credits are more likely to drop out. Learners who take a larger number of credits may then reduce as they understand which course they are more suited to or those taking a smaller number of credits may be more invested.

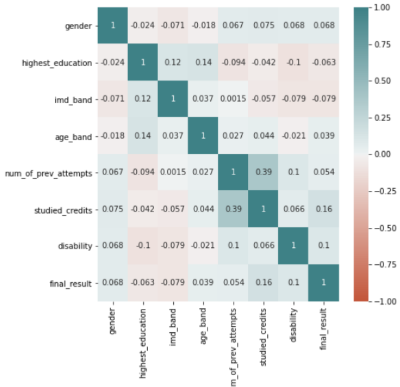

Figure 2: Pearson correlation coefficients for demographic features

The correlation coefficients for our extracted temporal features (Fig. 3) are much greater than the non-temporal features. We use the 25% proportion of the course duration feature vectors as part for demonstration of the analysis alongside the fact that features that are indicators from the earliest point in the course are the most valuable. Our generated features display a much higher level of correlation with the final result than the previous demographic data. This demonstrates the importance of temporal data as well as the potential for this features to capture the indication of high-dimensional features in a low dimensional space.

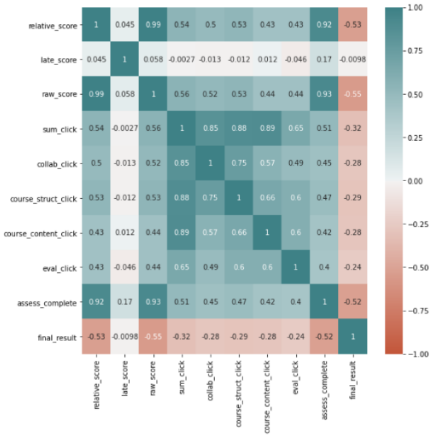

Figure 3: Pearson correlation coefficients for engineered temporal features

4.2 Predictive Models

Initial results from testing on each duration performance {25%, 50%, 75%, 100%}. As outlined previously we employ SMOTE to resample the dataset folds to produce a class-balanced training set each time. We take the average of the individual fold performances.

Table 1: Mean average F1 scores for classifiers trained upon varying data proportions

Classifier	Avg. F1 per proportion			
	25%	50%	75%	100%
Random Forest	**0.759**	**0.777**	0.780	**0.825**
ExtraTrees	0.758	0.765	0.782	0.807
Logistic Regression	0.714	0.749	0.771	0.793
SVM	0.366	0.366	0.366	0.366
Gradient Boosting	0.717	0.769	**0.787**	0.809

From initial testing we find that all models with the exception of SVM consistently improve as a larger proportion of the data is included, with a 9.8% increase from a proportion of 25% to 100% excluding SVM. Random Forest achieves the highest mean F1 score across three of the four proportions, being exceeded only by the performance of Gradient Boosting upon 75% of the temporal data.

Table 2: Random Forest mean average F1 scores for varying hyperparameters

Modification	Avg. F1 per proportion			
	25%	50%	75%	100%
Baseline	0.759	0.777	0.780	0.825
(*A*) Estimators=500	0.761	0.775	**0.782**	0.819
(*B*) Estimators=1000	0.763	0.777	0.781	0.818
(*C*) Criterion=Entropy	0.764	0.769	**0.782**	**0.827**
(*D*) Max Features= log_2	0.759	0.777	0.780	0.825
(*B, C*)	**0.766**	**0.781**	0.781	0.815

Selecting Random Forest as the highest performing ML technique we experimentally tune the model. We combine the two highest performing modifications of increasing the number of estimators to 1000 and changing the criterion from Gini to Entropy that measures the information gain rather than the probability of misclassification. Doing so improves upon the baseline performance of proportions {25%, 50%, 75%} by 0.5% on average at the cost of a decrease in the F1 score for {100%} but this is arguably the least valuable proportion for predictions.

Table 3 presents the results for our fine-tuned Random Forest predictive model. Considering the class imbalance, the models performs relatively well with a somewhat similar F1 score for each class suggesting that the SMOTE synthetic samples are from a similar if not the same distribution as the original samples. The class performance disparity is more pronounced with the 25% than 100% with *Withdrawn* having an F1 score 12.5% lower compared to 3.7% respectively. Model performance improvements as the data proportion increases is primarily due to improvements in *Withdrawn* classification, consistently improving unlike *Completed* classification that decreases for 50% and 75% before increasing at 100%.

4.3 Ensemble Model

Motivated by the performance of ensemble methods in prior work for this task [6], we evaluate an ensemble method composed of the three best performing models based upon Table 1; Random Forest, ExtraTrees, and Gradient Boosting. There are two types of voting within convential ensemble classifiers, hard voting that uses the predicted class labels, and soft that uses the greatest sum of the predicted class probabilities. We experiment with both and found soft voting to be slightly superior for this task.

We find that overall there is very little difference in performance between the individual Random Forest classifier and the ensemble classifier. The ensemble has a 0.3% lower average F1 score for {25%, 50%} and a 0.4% higher average F1 score for {75%, 100%}. The negligible performance difference suggests that the three classifiers in the ensemble have modelled very similar distributions and so there is no real gain in performance when combined. With these results the most effective solution is the individual Random Forest model that has the same performance without the added complexity and expensive of three different models.

5 DISCUSSION

Overall the Random Forest predictive model performs well upon the selected data. With a class imbalance of 60.1% to 39.9% the model average F1 scores of {0.766, 0.781, 0.781, 0.815}, far better than a model overfitting to the majority class and always predicting it which would produce an accuracy of 60%. There is still a slight bias towards predicting the class of *Completed* for a learner. For practical application it may be beneficial, if the model was going to have a bias, to have one towards predicting the *Withdrawn* class. This is because it may be favourable to ensure that all students who require intervention receive it and also provide support to some who don't rather than failing to intervene where necessary. There is a trade-off where such an approach would lead to a low precision that would make the system particularly inefficient for a teacher to use, as well as stretching resources unecessarily.

Our proposed solution performs similarly to prior work, [5] from which we have utilised features to capture higher level semantic concepts. [5] achieves an F1 score of 0.842 for a similar sized OULAD subset at 100% of the course data with Random Forest. At 0.833 our Random Forest solution is only 1.1% lower. Comparing to the results of [1] which, unlike [5], compared different course durations instead of only 100%, achieves {0.793, 0.841, 0.884, 0.907, 0.919} for the proportions of {20%,40%,60%,80%,100%} with Random Forest. Our results do fall slightly short but their work is not directly comparable due to considering a different problem. Rather than predicting dropout [1] predicts *Distinction/Pass* or *Fail/Withdrawn* and uses the entire dataset.

One limitation within this study is that there will be points at which the model is predicting whether a learner will dropout after the point that they have withdrawn. While this may not make complete semantic sense it does not degrade the quality of the model since the vectors for students after they have withdrawn remain the same and should still be classified as withdrawn. This approach is in line with prior work for this task [1, 4, 6] as well and enables the comparison with their performances.

Table 3: Performance of the experimentally tuned Random Forest

Proportion	F1 score			Precision	Recall	Accuracy
	Completed	Withdrawn	Average			
25%	0.817	0.715	0.766	0.769	0.764	0.779
50%	0.814	0.748	0.781	0.780	0.787	0.788
75%	0.808	0.754	0.781	0.781	0.791	0.787
100%	0.831	0.800	0.815	0.821	0.833	0.818

Table 4: Performance of the Ensemble Classifier

Proportion	F1 score			Precision	Recall	Accuracy
	Completed	Withdrawn	Average			
25%	0.814	0.716	0.765	0.769	0.763	0.777
50%	0.811	0.746	0.778	0.777	0.785	0.785
75%	0.814	0.758	0.786	0.786	0.796	0.792
100%	0.836	0.797	0.817	0.818	0.831	0.820

The ability of the model to learn to predict the occurrence of dropout, even with only 25% of the course data, using our engineered features demonstrates their ability to provide strong indicators for learners at risk of withdrawing while easily translating to higher level semantic concepts that can be leveraged by non-technical stakeholders. Further work could expand to use the entire dataset to directly compare to the results of [1] to evaluate the exact performance gain introduced by these further features alongside providing evidence for the generalisabilty of the solution. An alternate expansion of this work would be to investigate the ability for the same engineered features and models to be used in predicting when the learner will dropout, rather than just a binary classification of if they will at some point in the course. This would provide stakeholders with a more detailed timeline. For teachers this would enable them to organise support within the time available, rather than not knowing if the student would withdraw part way through the support they had begun offering. For course designers this would enable them to view at a per date level how their potential changes would affect dropouts and the time until dropout. This would offer more flexibility in the task of dropout reduction, enabling them to utilise techniques to increase the time until dropout but not necessarily prevent it in order to allow intervention and support to occur.

6 CONCLUSION

This work demonstrates that the engineered features of [1] and [5] provide strong indicators for likelihood of learners withdrawing from a particular module presentation. We find that these handcrafted features based upon temporal assessment and behavioural data are several times more strongly correlated on average with whether a leaner completes a course than non-temporal demographic data. Expanding upon the work of [1] we incorporate the semantic features of [5] to produce an effective predictive model for leaner dropout. After experimental fine-tuning our Random Forest model achieves an average F1 score of 0.766 on just 25% of the temporal data. Class wise the model achieves 0.817 and 0.715 for *Completed* and *Withdrawn* respectively after employed SMOTE augmentation, mitigating the possible model degradation due to the original 60.1% to 39.9% class imbalance. Future work will expand this approach to further subsets of the OULAD dataset in order to test the ability of the model to generalise as well as examine the possibility of adapting this approach to predict the date a learner will dropout, providing stakeholders with a more detailed view to facilitate effective intervention.

7 IMPLEMENTATION

https://colab.research.google.com/drive/18uKNY01zyuBmV2FEeM8Lab-x57fUp-5Z?usp=sharing

REFERENCES

[1] Muhammad Adnan, Asad Habib, Jawad Ashraf, Shafaq Mussadiq, Arsalan Ali Raza, Muhammad Abid, Maryam Bashir, and Sana Ullah Khan. 2021. Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access* 9 (2021), 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446

[2] Concepción Burgos, María L. Campanario, David de la Peña, Juan A. Lara, David Lizcano, and María A. Martínez. 2018. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers Electrical Engineering* 66 (2018), 541–556. https://doi.org/10.1016/j.compeleceng.2017.03.005

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (June 2002), 321–357. https://doi.org/10.1613/jair.953

[4] Evandro B. Costa, Baldoino Fonseca, Marcelo Almeida Santana, Fabrísia Ferreira de Araújo, and Joilson Rego. 2017. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior* 73 (2017), 247–256. https://doi.org/10.1016/j.chb.2017.01.047

[5] Fedia Hlioui, Nadia Aloui, and Faiez Gargouri. 2021. A Withdrawal Prediction Model of At-Risk Learners Based on Behavioural Indicators. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)* 16, 2 (March 2021), 32–53. https://ideas.repec.org/a/igg/jwltt0/v16y2021i2p32-53.html

[6] Sandeep Jayaprakash, Erik Moody, Eitel Lauria, James Regan, and Joshua Baron. 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics* 1 (05 2014), 6–47. https://doi.org/10.18608/jla.2014.11.3

[7] Suhang Jiang and Dimitrios Kotzias. 2016. Assessing the Use of Social Media in Massive Open Online Courses. arXiv:1608.05668 [cs.CY]

[8] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdráhal. 2017. Open University Learning Analytics dataset. *Scientific Data* 4 (11 2017), 170171. https://doi.org/10.1038/sdata.2017.171

[9] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos. 2009. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Comput. Educ.* 53 (2009), 950–965.

[10] J. R. Quinlan. 1986. Induction of Decision Trees. *Mach. Learn.* 1, 1 (March 1986), 81–106. https://doi.org/10.1023/A:1022643204877

[11] F. Wang and L. Chen. 2016. A Nonlinear State Space Model for Identifying At-Risk Students in Open Online Courses. In *EDM*.

[12] Philip Winne. 2010. Improving Measurements of Self-Regulated Learning. *Educational Psychologist* 45 (10 2010), 267–276. https://doi.org/10.1080/00461520.2010.517150