OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques

Nikhil Indrashekhar Jha[1], Ioana Ghergulescu[2][a] and Arghir-Nicolae Moldovan[1][b]
[1]*School of Computing, National College of Ireland, Dublin, Ireland*
[2]*Adaptemy, Dublin, Ireland*

Keywords: Dropout Prediction, MOOC, Learning Analytics, Machine Learning, OULAD Dataset.

Abstract: Massive Open Online Courses (MOOCs) have become increasingly popular since their start in the year 2008. Universities known worldwide for their traditional confined classroom education are also changing their practices by hosting MOOCs. These are Internet-based courses where students can learn at their own pace and follow their own schedule. Study materials and videos are provided that can be used in a blended learning program. Despite its many advantages, it suffers from problems such as high dropout and failure rates. Previous studies have mostly focused on predicting student dropout. This paper contributes to the body of research by investigating both student dropout and result prediction performance of machine learning models built based on different types of attributes such as demographic info, assessment info and interaction with the VLE. An analysis on the OULAD dataset showed that models based on student's interaction with the VLE achieved the high performance in terms of AUC, of up to 0.91 for dropout prediction and 0.93 for result prediction in case of Gradient Boosting Machine.

1 INTRODUCTION

Massive Open Online Courses (MOOCs) have been increasingly adopted and successful since their beginning in 2008. A course that was started by Sebastian Thrun and other members of the Stanford on Artificial Intelligence in the year 2011 attracted 1.6 million participants from more than 190 countries. According to Tan & Shao (2015) American universities or colleges that started some online course between 2008 and 2012, have maintained a steady growth of 10-20% every year. This has allowed them to reach the students that could not make it to college due to geographical boundaries or other limitations. Top universities such as MIT, Stanford and US Berkley have opened their online courses to a wide audience through MOOC platforms such as edX and Coursera. According to recent statistics by Class Central, in 2018 there were over 101 million students enrolled in more than 11400 online courses provided by over 900 universities (Shah, 2018). The number only increases with every passing day due to the self-paced learning scheme. Coursera is one the most popular MOOC platform with more than 37 million students joining courses.

Although MOOCs have become very popular in a short time span, they also experience problems with high student dropout and failure rates, as well as lack of support for struggling students. The free nature of MOOCs has resulted in many students dropping out from the course or not being able to obtain good grades. The dropout rate is typically 20% higher for students who are enrolled in online courses, while for some online courses from universities like Open University UK and China the dropout rate was even as high as 78% and 40% respectively (Tan & Shao, 2015). Xing & Du (2018) indicated that the dropout rate reached up to 90% for MOOC courses, which is considerably higher when compared to traditional campus courses.

This paper investigates ensemble methods, deep learning and regression techniques for predicting student dropout and final result in MOOCs. The analysis is done on the Open University Learning Analytics dataset (OULAD) (Kuzilek, Hlosta, & Zdrahal, 2017), a large complex dataset that requires

[a] https://orcid.org/0000-0003-3099-4221
[b] https://orcid.org/0000-0003-4151-1432

Jha, N., Ghergulescu, I. and Moldovan, A.
OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques.
DOI: 10.5220/0007767901540164
In *Proceedings of the 11th International Conference on Computer Supported Education (CSEDU 2019)*, pages 154-164
ISBN: 978-989-758-367-4

significant pre-processing and feature extraction that was not thoroughly investigated by previous research.

The rest of the paper is organized as follows. Section 2 presents related work on MOOC datasets and prediction. Section 3 presents the research methodology. Section 4 presents the evaluation results. Section 5 concludes the paper and presents future work directions.

2 RELATED WORK

Nowadays, many universities are using an analytical approach to help improve the online learning, teaching and assessment process. This section overviews some existing MOOC datasets and presents previous research works on MOOC learning analytics.

2.1 MOOC Datasets

Table 1 presents a summary of existing datasets from various MOOC platforms that were made available to researchers and used in different learning analytics research studies. Harvard and MIT are among main universities that are hosting various online courses using the edX platform and have made data available freely to enable learning analytics research. Several other universities such as George Mason University and NYU Stern School of Business have also published anonymised data to researcher for valuable knowledge discovery.

Open University UK is the largest academic institution in UK with around 170000 students enrolled in different programmes. Study materials related to the course are delivered through a VLE. The OULAD dataset contains data from 32,592 students and 22 module-presentations, and contains various data such as student demographic info, assessment dates and scores, and clickstream data of their interaction with the VLE. Other datasets such as such as KDD Cup and XuetangX only contain data related to the MOOC platform, but do not contain student demographic information.

2.2 MOOC Predictions

Table 2 provides a summary of previous research works that attempted to predict the result of the student at the end of the course (i.e. whether the student will pass or fail), or whether the student will drop out or complete the course. The table summarizes some of the machine learning models used by the researchers and relevant evaluation parameters provided by them.

Breslow et al. (2013) have devised a method that uses prior knowledge, skills, and activities such as the use of Virtual Learning Environment (VLE) and activities of the candidate to predict the end-of-MOOC performance of the candidate. Ashenafi, Riccardi, & Ronchetti (2015) have used data from the forum where students ask questions and rate answers. Data from the peer-assessment system was also used for predicting the result of the student.

Haiyang, Wang, Benachour, & Tubman, (2018) have used the OULAD dataset. The data contains records of 32593 students and the logs of their interaction with the Virtual Learning Environment (VLE). The clickstream data of the VLE environment was extracted on a daily basis and time series data frames were constructed for each of the modules. The VLE interaction log contains more than 10 million entries. The authors have used Time Series Forest to predict the dropout of students. Demographic information of the candidate was omitted from the model to avoid ethical issues. Open University (OU) provides resources like video lectures and some material that a student should read in course.

Hlosta, Zdrahal, & Zendulka (2017) have made use of the data about assignments that are submitted by the students. The authors point out that students who do not submit assignments have a 90% chance of dropping out from the course. The authors have also stressed on the previous educational background of the student. Haiyang et al. (2018) have extracted features form the VLE environment from the OULAD dataset, while Hlosta et al. (2017) has used existing features such as assignments, number of consecutive days the student is active, average median clicks and number of materials visited per day. The authors also point out the specificity problem of the OULAD dataset, which deals with the fact that the two groups of minority and majority students (i.e., submitting and not submitting assessments) change over time. Hong, Wei, & Yang (2017) pointed out that if we move back in time from the cut-off date of an assignment, the number of students who withdrew their enrolments outnumbered the students submitting assignments.

Alshabandar, Hussain, Keight, Laws, & Baker (2018) made used of assignments deadlines and the VLE database. The authors applied a probabilistic model which predicts at-risk students who may drop out from the course. For feature extraction, various VLE types are used for each student in some time interval and combined into a single value. Other features such as dynamic behavioural features,

Table 1: Summary of existing MOOC datasets.

Citation	Dataset / MOOC	Data Description
(Haiyang et al., 2018)	OULAD	32,592 students, over 10 million rows for VLE data, 22 different courses, data available as 7 CSV files
(Hlosta et al., 2017)		
(Alshabandar et al., 2018)		
(Piech et al., 2015)	Khan Academy	1.4 million exercises completed by 47495 students, 69 different types of exercises
(Hong et al., 2017)	XuetangX (OpenEdx)	course content, student enrolment, learning access log datasets used and combined them, 120,542 samples from 39 courses
(Xing & Du, 2018)	KDD Cup 2015	39 Courses, 7 csv data files
(Qiu et al., 2016)	XuetangX	11 Completed courses of Fall 2013 and Spring 2014. Computer Science and Electronics Engineering and non-science courses, 56,800,000 time-stamped records
(Al-Shabandar et al., 2017)	Dataset used from Harvard and MIT MOOC courses	597692 participants, 15 courses, 800,000-log file
(Liang, Li, & Zheng, 2016)	edX	39 courses
Cobos, Wilde, & Zaluska (2017)		8 in edX and 14 in FutureLearn, total of 22 courses, 8935 enrolled learners
(Balakrishnan & Coetzee, 2013)		29882 students, 4 assignments,4 exam grades data
(Chaplot, Rhim, & Kim, 2015)	Coursera	3 Million students, 5000 forum posts
(Tan & Shao, 2015)	Open University China	62375 students in three semesters of Fall 2010 and Fall 2011

demographic features, and different assignments feature were used. The prediction is made through Gaussian Finite Mixture model.

Heuer & Breiter, (2018) have also used time-based data of OULAD dataset for predicting whether the student will pass or fail a module. The authors pointed out that the daily activity of the student can be easily anonymized. According to the authors, a binary information about a student if he/she was active on a given day can be as meaningful as using the sensitive private data like gender, disability, and highest education. The authors have categorized students based on their daily use of VLE. All different types of activities of the VLE environment were combined into a single metric. The metric had dimensions equal to the number of days in a module. Each row denotes whether the student was active on the given day or not. The four target variables were then limited to two for binary classification as to whether the candidate will pass or fail. The authors then used k-means clustering technique to group data points which were similar. Different machine learning models were then used to predict the result of the candidate.

Kennedy, Coffrin, de Barba, & Corrin (2015) have used the previous educational background of the student as well as prior knowledge of the course and their engagement with the VLE to predict end-of-MOOC performance. The dataset was acquired from the Coursera platform and courses were hosted by the University of Melbourne. Different techniques such as knapsack and graph colouring were used. Different Knapsack points were derived which tested the prior knowledge of the participants. Graph colouring tested the problem solving related to computer science. Both were graded between 0 to 60. Apart from these different assignments, the number of days taken to complete the assignment and total points scored were used as parameters to prepare the model.

Dalipi, Imran, & Kastrati (2018) in their review paper have mentioned that machine learning is the easy part of the process. The key and the difficult part is the feature selection from the huge amount of data that is made available by the MOOC platform. Different platforms offer data in different ways which makes it sparse and thus has a lot of features. The authors have studied two different MOOC datasets and categorized them into two parts based on the types of features available. One is student related and the other is MOOC related. In the student related dataset, the features are generally related to the behavioural aspect of the student while the MOOC related dataset has features that describe the course and the modules. The authors also argue that even though a lot of data is available, it still lacks features that can accurately point out not only dropout students but also the result of the candidates.

Table 2: Summary of MOOC prediction research studies.

Citation	Algorithms	AUC	Accuracy	Train Split	Test Split	Class Values
(Haiyang et al., 2018)	TSF	-	0.93	-	-	Dropout / No Dropout
(Hlosta et al., 2017)	SVM	0.779	-	-	-	Dropout / No Dropout
	NB	0.678	-			
	XGB	0.744	-			
	LR	0.756	-			
(Alshabandar et al., 2018)	EDDA	0.954	0.920	-	-	Dropout / No Dropout
	KNN	0.951	0.911			
	LR	0.949	0.893			
(Hong et al., 2017)	SVM	0.795	0.861	80%	20%	Dropout / No Dropout
	RF	0.852	0.865			
	MLR	0.855	0.861			
	SVM-C	0.909	0.904			
	C-RF	0.932	0.927			
	C-MLR	0.916	0.910			
(Xing & Du, 2018)	KNN	0.947	0.966	70%	30%	Dropout / No Dropout
	SVM	0.976	0.944			
	DT	0.876	0.982			
	DL	0.984	0.974			
(Li et al., 2016)	KNN	0.947	0.966	80%	20%	Certificate Attainment and Grade Prediction
	SVM	0.976	0.944			
	DT	0.876	0.982			
	DL	0.984	0.974			
(Qiu et al., 2016)	LRC	0.809	-	From 20% to 90%	From 80% to 10%	Dropout / No Dropout
	SVM	0.709	-			
	FM	0.903	-			
	LadFG	0.962	-			
(Al-Shabandar et al., 2017)	DT	0.998	0.973	70%	30%	Grades Prediction
	RF	0.997	0.985			
	SVM	0.991	0.973			
	SOM	0.956	0.956			
	NB	0.987	0.962			
	NN	0.994	0.972			
	LR	0.988	0.954			
	LDA	0.986	0.954			
(Liang et al., 2016)	GBDT	-	0.879	60%	40%	Dropout / No Dropout
(Cobos et al., 2017)	GBM	-	0.83	-	-	Dropout / No Dropout
	kNN	-	0.79			
	LogitBoost	-	0.81			
	XGB	-	0.82			
(Tan & Shao, 2015)	ANN	-	0.940	70%	30%	Dropout / No Dropout
	DT	-	0.946			
	BN	-	0.940			

Legend: **BKT** Bayesian Knowledge Tracing, **BLR** Boosted Logistic Regression, **DL** Deep Learning, **DRSA** Dominance Based Rough Set Approach, **DT** Decision Tree, **EDDA** - Eigenvalue Decomposition Discriminant Analysis**, GBRM** Generalized Boosted Regression Model, **GDBT** Gradient Boosting Decision Tree, **HMM** Hidden Markov Model, **KNN** K-nearest Neighbour, **LadFG** Latent Dynamic Factor Graph, **LDA** Linear Discriminant Analysis, **LR** Logistic Regression, **LSTM** Long Short Term Memory, **NB** Naïve Bayes, **NN** Neural Network, **RF** Random Forest, **RNN** Recurrent Neural Network, **SOM** Self-Organized Map, **SVM** Support Vector Machine, **TSF** Time Series Forest, **WKNN** Weighted K-Nearest Neighbour, **XGB** Extreme Gradient Boosting.

To overcome this Nagrecha, Dillon, & Chawla, (2017) has used certain VLE features that are extracted while the student was watching the course videos. A pattern such as play, pause, rewind, forward is observed to generate clickstream data. These VLE recorded data was then used to extract patterns and through visualizations, dropout of the student was predicted on a weekly basis.

Liu & Li (2017) have taken a hybrid approach where they used clusters to make groups of active and non-active participants using the k-means algorithm. Association rules are then discovered using the Apriori algorithm. The rules discovered were then used to predict the dropouts from the course with 80% accuracy.

Wang, Yu, & Miao (2017) have also built a hybrid model that combines Convolution Neural Network and Recurrent Neural Network to predict the outcome of the student in the course. The model is divided in two layers. The bottom layer does the job of convolution and pooling to extract the features in different time periods. The upper layer then receives the selected features and combines the information for prediction.

Boyer & Veeramachaneni (2015) categorised students based on the courses. The authors pointed out that the behaviour of the participants that were enrolled in a course in the past can be used to predict if they will drop out in the future if they participate in the same course as before or will go on to complete the course. It estimates the distribution of the weights beforehand by splitting data in 10 samples prepared randomly from the main dataset. Then for each sample, Logistic Regression is applied to predict the drop out of the student.

Chaplot et al. (2015) in his work has used clickstream data from the Coursera platform. The data consists of three million students VLE data from 5000 forum posts. Most of the features related to students' sentiments were used for analysis. Sentiment analysis from the student forum platform was used to carry out the sentiment analysis. The score of these sentiments was extracted. The sentiment score was marked 1 if the student is predicted to drop out from the course or 0 otherwise. The model was built using Artificial Neural Networks. The downside was that the model cannot be interpreted as it is a black box method. Moderate results were obtained as the data was highly imbalanced.

2.3 Research Gaps

In this research work, we decided to use the OULAD dataset provided by the Open University UK. Thus, we keep this discussion to the papers that used this dataset in their work. Haiyang et al. (2018) in his work has made use of time series forest, based on three categorical values of the activity_type attribute from the studentVle dataset. However, there are 17 more different values for that attribute that represent different resources available. These resources were not used in their work. It also does not use any data from the assessment dataset.

Hlosta et al. (2017) have used both demographic information and student interaction with the VLE which is also used in our work. However, their work does not use the assessment scores, and instead they use whether the student has submitted the assessment or not. From the data, it can be inferred that there are 18042 students who have not submitted at least one assessment and did not dropout from the course. Similarly, Alshabandar et al. (2018) have also considered behavioural data of the students and submitted assessments but not their score.

This research work complements previous studies by investigating the performance of ensemble, deep learning and regression techniques to predict student dropout and result based on different groups of attributes, specifically demographic info, assessment scores and VLE interaction information.

3 METHODOLOGY

This research followed the Knowledge Discovery in Databases (KDD) methodology.

3.1 Dataset Selection and Description

Table 1 presented a summary of datasets reviewed during the data selection step. Following the review, the decision was to use the OULA[1] dataset from Open University UK. According to Kuzilek et al. (2017) the type of information can be divided in three parts: demographic, assessment, and VLE interaction. It includes the result of the assessments submitted by the students, while the detailed VLE clickstream data enables researchers to engineer various features and build various models for predicting students' performance during the course.

[1] https://analyse.kmi.open.ac.uk/open_dataset

Table 3: Summary of the OULAD dataset before feature engineering.

Data File	No of Records	Description	Attributes
courses	22	Information about the courses	code_module, code_presentation, module_presentation_length
studentInfo	32593	Contains demographic information about the student	code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, final_result
studentRegistration	32593	Registration of the student for a course presentation	code_module, code_presentation, id_student, date_registration, date_unregistration
assessments	196	Assessments for every course presentation	code_module, code_presentation, id_assessment, assessment_type, date, weight
studentAssessments	173740	Assessments submitted by the students	id_assessment, id_student, date_submitted, is_banked, score
vle	6365	Online learning resources and materials	id_site, code_module, code_presentation, activity_type, week_from, week_to
studentVle	1048575	Student interaction with the VLE resources	code_module, code_presentation, id_student, id_site, date, sum_click

Table 4: Summary of the dataset after feature engineering.

Attributes Category	Number Attributes	Attributes	Type
Student Registration Information	3	Student_Id, Code_module, CodePresentation	Categorical
Student Demographic Information	6	Gender, Region, Highest_Education, IMD_Band, Age_Band, Diability	Categorical
	2	Studied_Credits, Number_of_Previous_Attempts	Numeric
Assessments Information	3	Avg_TMA_Score, Avg_CMA_Score, Exam_Score	Numeric
Sum of clicks for each VLE activity_type	20	Sum_Clicks_{resources, oucontent, url, homepage, subpage, glossary, forumng, oucollaborate, dataplus, quiz, ouelluminate, sharedsubpage, questionnaire, page, externalquiz, ouwiki, dualpane, repeatactivity, folder, htmlactivity}	Numeric
Number of visits for each VLE activity_type	20	Count_Visits_{resources, oucontent, …, htmlactivity}	Numeric

Table 3 provides a summary of the OULAD dataset. The dataset comes as 7 separate csv files and requires significant processing and transformation to extract features before building prediction models. A student demographic information is linked with other information segments of assessments and VLE interactions.

3.2 Pre-processing and Transformation

This subsection details the data pre-processing and transformation carried out before building predictive models. Table 4 provides a summary of the dataset after the feature extraction. Features from different files were extracted based on three main keys that identify a student uniquely in the entire database: id_student, course_module, module_presentation.

The assessments data consisted of different types of assessments submitted by students: teacher marked assessments (TMA), computer marked assessments (CMA) and Exam. However, only the average assessment scores were used as predictors, while the Exam results were excluded from the analysis.

The VLE data has information about 20 different types of online resources. New features such as the sum of clicks and number of visits were created for each unique resource type based on the studentVle data. The approach is different of previous works. Haiyang et al. (2018) aggregated the clickstream data for each day of the course, thus if the course was over 279 days they added 279 new features for every student. Alshabandar et al. (2018) have also used the sum of clicks but not for all resource types.

3.3 Data Mining and Evaluation

This research work aims to investigate the performance of the student in the course. The research question was then broken down in two parts:

- to predict whether a student will drop out from the course.
- to predict whether a student that does not drop out will pass or fail the course.

The following machine learning classification algorithms were considered for predicting the dropout and final result:

- Distributed Random Forest (DRF) is one of the powerful classifications and regression alghoritms that can be used for both multi and binary classification types accurately. It uses an ensemble method for learning that generates a multitude of trees that is collectively referred to as forest. Each of the trees in the forest is a weak learner that is built on a different part of the dataset provided. By combining the strengths of different trees, the random forest algorithm can achieve increased performance.
- Gradient Boosting Machine (GBM) is used for regression and classification problems. It also uses an ensemble method to predict the outcome of the variable. For classification problems it makes use of logarithm loss function. Its strength lies in making use of the residual patterns and use them to train itself. These patterns in residuals strengthen models with weak predictions.
- Deep Learning (DL) is a method of training artificial neural networks with more than two non-output layers. It implements techniques for learning data representations and create new features from raw data which replaces in part the need for feature engineering. The Deep Learning model with feedforward neural network that is trained with gradient descent using back-propagation was used.
- Generalized Linear Model (GLM) is a generalization of other models that includes linear regression, logistic regression, ANOVA, Poisson regression, ordinal regression, log-linear models, etc. It assumes that the outcome variable follows an exponential distribution. GLM can be used for different regression and classification problems.

Two approaches were used for evaluating the performance of the machine learning models. The first approach was 10-fold cross validation, while the second approach was holdout method with 75% of data used for training and 20% for validation. The AUC (Area under the Receiver Operating Characteristics Curve) was used to measure the model performance. AUC captures the ability of a binary classifier to separate between classes as the discrimination threshold is varied, and as such it is a more comprehensive metric as compared to accuracy that can take different value at different threshold.

3.4 Implementation

Figure 1 illustrates the methodology from an implementation perspective. The original dataset was retrieved as one compressed zip from the OULAD. The 7 csv files that comprise the raw data were brought into the staging area of the database. SQL Server [2] was used for data pre-processing and transformation due to the vle file containing more than 10 million records. Some additional tables were created to store intermediate transformations. Data pre-processing, transformation, and final dataset preparation were conducted by making use of SQL Server Management Studio (SSMS) and SQL Server Integration Service (SSIS).

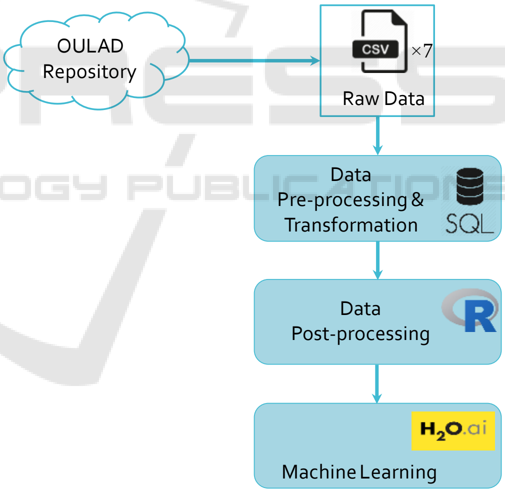

Figure 1: Implementation architecture.

The transformed dataset was imported in R[3] for further post-processing such as feature selection. The post-processed data is then imported in the H20[4] framework for training and validation of the machine learning models. The performance data is extracted and processed for visualization and presentation.

[2] https://www.microsoft.com/en-us/sql-server/

[3] https://www.r-project.org/

[4] https://www.h2o.ai/

4 RESULTS

This paper investigates the performance of various machine learning algorithms to predict whether a student will dropout from the course, and for students that do not dropout weather they pass or fail the course. Table 5 summarises the number of instances used for the two cases.

Table 5: Number of instances.

Classification Type	Train / Cross Validation Instances	Validation Instances
Dropout	24445	8149
Result	16828	5609

Four different experiments were conducted for both dropout and result classification to assess the prediction performance of machine learning models built based on different categories of predictors, specifically: demographic info, assessments scores, VLE interactions, and all attributes. The id_student, code_module, module_presentation and exam_score were excluded and not used as predictors.

Table 6: Summary of evaluation experiments used for dropout and final result prediction.

Experiment	Predictors Category	Number Predictor Attributes
1	Demographics	8
2	Assessment Scores	2
3	VLE Interactions	40
4	All Attributes	50

4.1 Dropout Prediction

Figure 2 presents the AUC performance results for the dropout classification models. The results show that the models created based on demographics information achieved between 0.61 and 0.64 AUC on the validation set. GBM and DL models have slightly higher performance, but they tend to overfit to the training data and more hyperparameter tuning would be required to optimise them. As compared GLM and DRF models had slightly lower but more stable performance and were faster to train.

The models created based on the assessment scores achieved over 0.82 AUC, and as high as 0.84 for GBM. The models offered stable performance for training, validation and cross-validation.

The models based on VLE interaction features achieved around 0.88 AUC for GLM, and 0.90 for DL, GBM and DRF on the validation data.

The results also show that the models based on all attributes (i.e., demographic, assessments and VLE interactions), only achieved about 0.01 higher AUC than the models based on the VLE interactions only.

Comparing the results to those of previous research works that used the OULAD dataset, our results are better than the results achieved by (Hlosta et al., 2017), but lower than those achieved by (Alshabandar et al., 2018). However, both of those previous works have used temporal features and looked at the performance of the algorithms on various days or intervals.

4.2 Result Prediction

Figure 2 also presents the AUC performance results for the result classification models. The results show that the models created based on demographics information achieved between 0.62 and 0.65 AUC on the validation data. The GLM model had the highest and stable performance, while the GBM and DL models achieved the highest training AUC but would require further optimisation.

The models created based on the assessment scores achieved an AUC performance ranging from 0.79 in case of DRF and 0.82 for GBM, on the validation set.

The models based on VLE interaction features achieved around 0.90 AUC on the validation data and offer stable performance. Moreover, the models based on all attributes, only achieved about 0.01 higher AUC than the models based on the VLE interactions alone.

5 CONCLUSIONS

While MOOCs have become increasingly popular over the past decade they are facing high dropout and failure rates. Most previous research have focused on dropout prediction. This research investigated the performance of ensemble, deep learning and regression techniques to predict student dropout, as well as if students that do not drop out will pass or fail the course. Various models were built on different categories of attributes (i.e., demographic info, assessment scores, and VLE interaction data). The analysis was conducted on the recent OULAD dataset that was not thoroughly investigated. The results showed that machine learning models based on student’s interaction with the VLE achieved high

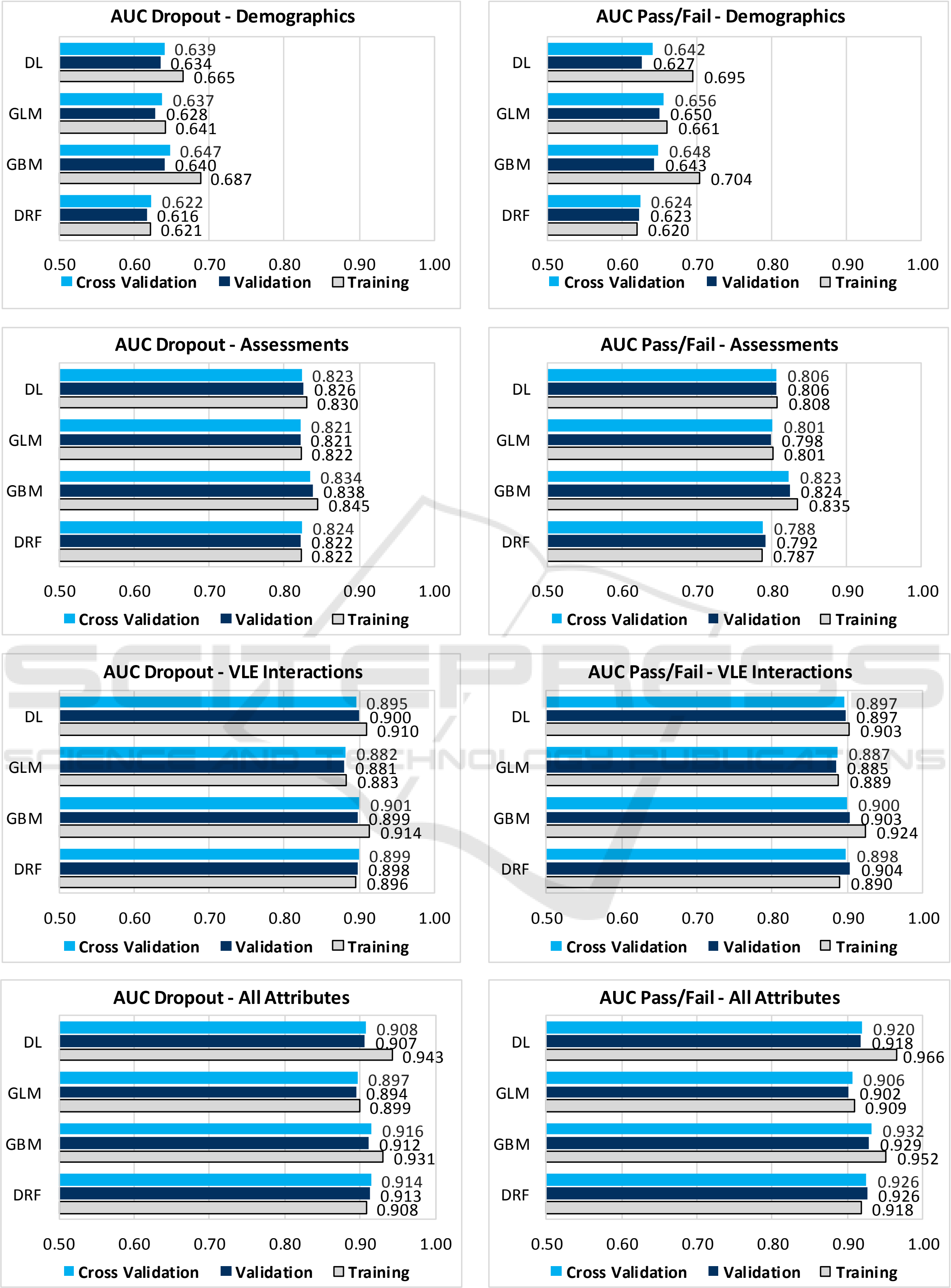

Figure 2: AUC performance results for binary dropout (yes/no) and result (pass/fail) classification models.

performance in terms of AUC, of up to 0.91 for dropout prediction and 0.93 for result prediction in case of Gradient Boosting Machine. The results also showed that considering student demographics info and assessment scores along with the VLE interactions leads to a small 0.01 increase in AUC.

This research has focused on aggregate features and did not made use of the date attribute available for student assessments and VLE interactions. Future work directions would focus on feature selection and engineering, including time based metrics related to assessments and student interactions to improve the dropout and result prediction performance.

REFERENCES

Alshabandar, R., Hussain, A., Keight, R., Laws, A., and Baker, T. (2018). The Application of Gaussian Mixture Models for the Identification of At-Risk Learners in Massive Open Online Courses. In *2018 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1–8). https://doi.org/10.1109/CEC.2018.8477770

Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., and Radi, N. (2017). Machine learning approaches to predict learning outcomes in Massive open online courses. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 713–720). https://doi.org/10.1109/IJCNN.2017.7965922

Ashenafi, M. M., Riccardi, G., and Ronchetti, M. (2015). Predicting students' final exam scores from their course activities. In *2015 IEEE Frontiers in Education Conference (FIE)* (pp. 1–9). https://doi.org/10.1109/FIE.2015.7344081

Balakrishnan, G., and Coetzee, D. (2013). *Predicting student retention in massive open online courses using hidden markov models* (No. UCB/EECS-2013-109). Electrical Engineering and Computer Sciences University of California at Berkeley. Retrieved from https://www2.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-109.pdf

Boyer, S., and Veeramachaneni, K. (2015). Transfer Learning for Predictive Models in Massive Open Online Courses. In C. Conati, N. Heffernan, A. Mitrovic, and M. F. Verdejo (Eds.), *Artificial Intelligence in Education (AIED 2015)* (pp. 54–63). Springer International Publishing.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., and Seaton, D. T. (2013). Studying Learning in the Worldwide Classroom Research into edX's First MOOC. *Research & Practice in Assessment*, *8*, 13–25.

Chaplot, D. S., Rhim, E., and Kim, J. (2015). Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks. In *Proceedings of the Workshops at the 17th International Conference on Artificial Intelligence in Education AIED 2015; Volume 3: Fourth Workshop on Intelligent Support for Learning in Groups (ISLG)* (pp. 7–12). Madrid, Spain. Retrieved from http://ceur-ws.org/Vol-1432/islg_pap2.pdf

Cobos, R., Wilde, A., and Zaluska, E. (2017). Predicting attrition from massive open online courses in FutureLearn and edX. In *Joint MOOCs workshops from the Learning Analytics and Knowledge (LAK) Conference 2017* (pp. 74–93). Simon Fraser University, Vancouver, BC, Canada. Retrieved from http://ceur-ws.org/Vol-1967/FLMOOCS_Paper5.pdf

Dalipi, F., Imran, A. S., and Kastrati, Z. (2018). MOOC dropout prediction using machine learning techniques: Review and research challenges. In *2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1007–1014). https://doi.org/10.1109/EDUCON.2018.8363340

Haiyang, L., Wang, Z., Benachour, P., and Tubman, P. (2018). A Time Series Classification Method for Behaviour-Based Dropout Prediction. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)* (pp. 191–195). https://doi.org/10.1109/ICALT.2018.00052

Heuer, H., and Breiter, A. (2018). Student Success Prediction and the Trade-Off between Big Data and Data Minimization. In D. Krömker and U. Schroeder (Eds.), *DeLFI 2018 - Die 16. E-Learning Fachtagung Informatik*. Bonn, Germany: Gesellschaft für Informatik e.V. Retrieved from http://dl.gi.de/handle/20.500.12116/16955

Hlosta, M., Zdrahal, Z., and Zendulka, J. (2017). Ouroboros: Early Identification of At-risk Students Without Models Based on Legacy Data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 6–15). New York, NY, USA: ACM. https://doi.org/10.1145/3027385.3027449

Hong, B., Wei, Z., and Yang, Y. (2017). Discovering learning behavior patterns to predict dropout in MOOC. In *2017 12th International Conference on Computer Science and Education (ICCSE)* (pp. 700–704). Houston, TX, USA. https://doi.org/10.1109/ICCSE.2017.8085583

Kennedy, G., Coffrin, C., de Barba, P., and Corrin, L. (2015). Predicting Success: How Learners' Prior Knowledge, Skills and Activities Predict MOOC Performance. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 136–140). New York, NY, USA: ACM. https://doi.org/10.1145/ 2723576.2723593

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

Li, W., Gao, M., Li, H., Xiong, Q., Wen, J., and Wu, Z. (2016). Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning. In *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 3130–3137). Vancouver, BC, Canada. https://doi.org/10.1109/IJCNN.2016.7727598

Liang, J., Li, C., and Zheng, L. (2016). Machine learning application in MOOCs: Dropout prediction. In *2016 11th International Conference on Computer Science*

Education (ICCSE) (pp. 52–57). Nagoya, Japan. https://doi.org/10.1109/ICCSE.2016.7581554

Liu, T., and Li, X. (2017). Finding out Reasons for Low Completion in MOOC Environment: An Explicable Approach Using Hybrid Data Mining Methods. In *2017 International Conference on Modern Education and Information Technology (MEIT 2017)* (pp. 376–384). Chongqing, China. https://doi.org/10.12783/dtssehs/meit2017/12893

Nagrecha, S., Dillon, J. Z., and Chawla, N. V. (2017). MOOC Dropout Prediction: Lessons Learned from Making Pipelines Interpretable. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 351–359). Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3041021.3054162

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., and Sohl-Dickstein, J. (2015). Deep Knowledge Tracing. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 505–513). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5654-deep-knowledge-tracing.pdf

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q., and Xue, Y. (2016). Modeling and Predicting Learning Behavior in MOOCs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 93–102). New York, NY, USA: ACM. https://doi.org/10.1145/2835776.2835842

Shah, D. (2018, December 11). By The Numbers: MOOCs in 2018 - Class Central. Retrieved March 18, 2019, from https://www.class-central.com/report/mooc-stats-2018/

Tan, M., and Shao, P. (2015). Prediction of student dropout in e-learning program through the use of machine learning method. *International Journal of Emerging Technologies in Learning (IJET)*, *10*(1), 11–17.

Wang, W., Yu, H., and Miao, C. (2017). Deep Model for Dropout Prediction in MOOCs. In *Proceedings of the 2Nd International Conference on Crowd Science and Engineering* (pp. 26–32). New York, NY, USA: ACM. https://doi.org/10.1145/3126973.3126990

Xing, W., and Du, D. (2018). Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention. *Journal of Educational Computing Research*, 0735633118757015. https://doi.org/10.1177/07356331 18757015