RESEARCH ARTICLE OPEN ACCESS

Machine Learning Techniques to Predict Dropout Prediction of MOOCs

Kinjal Patel[1], Kiran Amin[2]
1Doctoral Student, Ganpat University
Email: cekinjalvp@gmail.com)
2 Pro-Vice Chancellor and Executive Dean, Ganpat University
Email: kiran.amin@ganpatuniversity.ac.in

Abstract:

MOOCs are very popular all over the world. There are still limitations that need to be addressed. One of the major issues is the high dropout rate. This focuses on dropout rate prediction using a range of machine learning techniques. In this paper. the The Open University Learning Analytics Dataset (OULAD) is used to train and test Decision Tree, Random Forest, Gaussian naive bayes, AdaBoost classifier, Extra tree classifier, XGBoost classifier, Multilayer Perceptron (MLP) models techniques to predict student dropout, as well as if students that do not drop out will pass or fail the course. The XGBoost classifier performs better than others with an accuracy of 89% in forecasting whether a student will pass or fail. At the same time, the XGBoost has the highest accuracy of 94% when predicting whether a student will drop out or continue the course.

Keywords —MOOCS, drop-out prediction, OULAD, Machine learning,

I. INTRODUCTION

With the advancement of technology, educational institutions are moving towards providing access to scalable e-learning solutions that allow learners to access content on any device and at any time, learn at their own pace, and customize their curriculum with industry-tailored courses [1]. This has led to the emergence of Massive Open Online Courses (MOOCs). MOOCs are online courses that are designed for large-scale participation and are typically available to anyone with internet access [2]. They offer a variety of subjects and are often free or low-cost, making education more accessible to individuals worldwide [3]. In addition, MOOCs can provide advantages for colleges and universities seeking to enhance their distance learning offerings [1].

MOOCs were first introduced in 2008 by Georges Siemens and have revolutionized distance education due to their openness, simplicity, quality, and massive reach [4]. The year 2012, also known as "The Year of the MOOC," saw a surge in the popularity of MOOCs with the launch of various platforms such as Coursera, edX, and Udacity [5].. These courses offer free access to subject areas from top universities worldwide and enable learners to interact with professors and peers across the globe for help and support [3]. MOOCs provide the flexibility to learn at any time and from any location, making them valuable for both learners and trainers who offer free or paid certifications online [4, 6].

Massive Open Online Courses (MOOCs) have gained significant enrollments, especially among younger learners who have easy access to online resources. The modern MOOC movement has grown rapidly, with the top five providers such as Coursera, edX, FutureLearn, Udacity, and Swayam reaching more than 110 million learners in 2019 (source: Class Central). The COVID-19 pandemic has further increased the demand for online education, leading to significant growth for MOOC providers. In April 2020, the top three MOOC

providers (Coursera, edX, and FutureLearn) registered as many new users as they did in the entire year of 2019 (source: Center).

Despite the widespread popularity of Massive Open Online Courses (MOOCs), there are still limitations that need to be addressed. One of the major issues is the high dropout rate, with less than 10% of enrollees completing the course and receiving a certificate [7]. This low completion rate is evidenced by several examples, such as a software engineering course offered by MIT and Berkeley, which had 50,000 registrations but only a 7% pass rate [6]. Duke University's Bioelectricity MOOC received 12,175 registrations, but only 2.6% completed the course [8]. Improving enrollment, completion, and overall learner experience is crucial to address these challenges. One solution is to develop efficient student success prediction models that can predict student dropout, completion, and learning in MOOCs [9]. Once an effective predictive model is found, personalized interventions can be implemented to improve learner outcomes and streamline the interaction between learners and instructors [7, 9].

This paper describes a comprehensive investigation of dropout rate prediction using a range of machine learning techniques. For this purpose, we used "The Open University Learning Analytics Dataset (OULAD)" to train and test the models.

II. LITERATURE REVIEW

With the increasing popularity of MOOCs, a large number of individuals enroll in these courses, and their activities are recorded in log files by MOOC providers [10]. This has resulted in the creation of numerous datasets, and many researchers are exploring ways to analyze this data and extract valuable insights from it. In the following paragraphs, we will provide a concise summary of the different approaches and models created by several researchers to predict the dropout rates of MOOC learners using diverse datasets.

Several studies have been conducted to predict MOOC student dropouts using different methods. For instance, Taylor, Veeramachaneni [11] used clickstream and forum submission data to train a logistic regression classifier to predict whether students would stop learning. He, Bailey [12] also used logistic regression to predict dropout based on course completion, assignment completion, and scoring. Another approach involves using time series classification methods, such as hidden Markov chains, nonlinear state space models, and RNNs. Kizilcec, Piech [13] identified four classes of learner engagement within MOOCs based on video lecture and assignment grades, and used clustering techniques to describe engagement activity. Mubarak, Cao [14] used a predictive model that combines logistic regression with an input-output hidden Markov model. Costa, Fonseca [15] implemented four ML algorithms to identify students at risk of failure, with the Support Vector Machine (SVM) achieving the highest accuracy. Finally, Baker, Evans [16] analyzed clickstream data to explore the relationship between online engagement and academic performance.

The utilization of predictive models discussed in this paper may assist educational institutions and instructors in promptly detecting students who are at risk of academic struggles, thus enabling them to intervene and provide suitable persuasive techniques to motivate these students to improve their performance and stay on track.

III. METHODOLOGY

We used The Open University Learning Analytics Dataset (OULAD) to train and test the models to investigation of dropout rate prediction using a range of machine learning techniques in this study. The dataset was published by Open University, which is a publicly funded British university. The university has the highest number of undergraduate students in the United Kingdom. Established in 1969, it is also the largest academic institution in the UK, and one of the largest in Europe, with a total enrollment of 2 million

students. As implied by its name, Open University primarily serves off-campus students.

The OULAD contains over 300,000 records of student activity, including log data from virtual learning environments (VLE), information about seven courses offered, student demographic information, and course-related data such as grades and assessments. The dataset was collected as part of the Open University Learning Analytics project, which aims to improve student success and retention through the use of analytics and data-driven interventions.

IV. DATA PROCESSING

Preprocessing is indeed the process of transforming raw data into a useful dataset. Preprocessing also includes a range of techniques such as data cleaning, normalization, scaling, feature extraction, and feature selection, among others. Data cleaning involves identifying and correcting errors in the dataset, such as incorrect data types, outliers, and inconsistencies. Normalization and scaling are techniques used to transform the data to a common scale, so that different features can be compared and analyzed more easily. Feature extraction involves identifying and selecting the most relevant features or variables from the dataset, while feature selection aims to identify and remove redundant or irrelevant features. Preprocessing It is a critical step and requires careful consideration and expertise to ensure that the resulting dataset is accurate, relevant, and useful for the intended purpose. In this paper, we focused on dealing with missing data through two methods: removing records that were incomplete or filling in the gaps with mean or mode values, which was determined based on the type of data being analyzed. For example, for the studentAssessment file, incomplete records with only a few missing values were removed. In the student registration file, the missing values in the data_registration column were replaced with zeros. As for the studentInfo file, the imd_band column contained categorical missing data, so the mode was used to fill in those gaps.

The Vle file contains multiple types of activities, including resource, subpage, oucontent, and url, among others. In order to analyze the data, I generated a count plot for each activity type. This visualization showed that the 'resource' activity type had the highest number of data points in the Vle file.

After reviewing the 'date_submitted' and 'date' columns, three new attributes were created: 'click_timing', 'before_click', and 'after_click.' These attributes helped to determine whether students submitted their assignments on time or late, by comparing the submission date to the deadline. An 'on-time submission' was assigned a value of 1, while a 'late submission' was assigned a value of 0. Then, various data files were merged as required and encoded categorical variables as either nominal (with no category order) or ordinal (with a category order).

A correlation matrix and heat map was utilized to identify the most significant factors for predicting student performance by analyzing the relationship between different features. We were able to identify any interdependencies by examining the correlation between independent and dependent variables. The correlation matrix revealed that the number of clicks and assessment score had the most significant impact on predicting student outcomes and dropout rates by displaying the interrelationships between various features.

TABLE I
CORRELATION ANALYSIS

	highest_education	num_of_prev_attempts	studied_credits	sum_click	After_Clicks	Before_Clicks	date_registration	module_presentation_length	date_submitted	is_banked	score	date	weight	Result	dropout
highest_education	1	-0.0239	0.01096	0.06038	0.05956	0.04338	-0.06	0.00819	-0.0009	0.00471	0.05905	-0.021	0.05536	-0.0541	-0.0132
num_of_prev_attempts	-0.0239	1	0.22497	-0.0505	-0.0495	-0.04	0.04082	-0.0629	-0.0599	0.29079	-0.0664	-0.0342	-0.0106	0.10973	0.03783
studied_credits	0.01096	0.22497	1	0.06328	0.06201	0.0507	0.06335	-0.0335	-0.0569	0.04331	-0.0476	-0.0481	0.03268	0.06005	0.05945
sum_click	0.06038	-0.0505	0.06328	1	0.99792	0.57193	0.04798	0.06608	0.06664	-0.067	0.18846	0.21265	-0.0231	-0.2517	-0.156
After_Clicks	0.05956	-0.0495	0.06201	0.99792	1	0.51779	0.04391	0.06812	0.07105	-0.0667	0.18825	0.21623	-0.0258	-0.255	-0.1599
Before_Clicks	0.04338	-0.04	0.0507	0.57193	0.51779	1	0.0779	0.01012	-0.0197	-0.0407	0.10552	0.07051	0.021	-0.0953	-0.0351
date_registration	-0.06	0.04082	0.06335	0.04798	0.04391	0.0779	1	0.06195	-0.0282	0.01244	-0.0184	-0.0174	0.04392	-0.0085	0.02071
module_presentation_length	0.00819	-0.0629	-0.0335	0.06608	0.06812	0.01012	0.06195	1	0.04782	0.02846	0.01433	0.08333	0.05247	-0.053	-0.0134
date_submitted	-0.0009	-0.0599	-0.0569	0.06664	0.07105	-0.0197	-0.0282	0.04782	1	-0.1725	-0.0339	0.79715	0.23809	-0.2153	-0.2219
is_banked	0.00471	0.29079	0.04331	-0.067	-0.0667	-0.0407	0.01244	0.02846	-0.1725	1	-0.0081	-0.0612	-0.0146	0.10487	0.0604
score	0.05905	-0.0664	-0.0476	0.18846	0.18825	0.10552	-0.0184	0.01433	-0.0339	-0.0081	1	0.07606	-0.1664	-0.3177	-0.1471
date	-0.021	-0.0342	-0.0481	0.21265	0.21623	0.07051	-0.0174	0.08333	0.79715	-0.0612	0.07606	1	-0.0127	-0.2277	-0.2158
weight	0.05536	-0.0106	0.03268	-0.0231	-0.0258	0.021	0.04392	0.05247	0.23809	-0.0146	-0.1664	-0.0127	1	-0.0436	-0.0508
Result	-0.0541	0.10973	0.06005	-0.2517	-0.255	-0.0953	-0.0085	-0.053	-0.2153	0.10487	-0.3177	-0.2277	-0.0436	1	0.50877
dropout	-0.0132	0.03783	0.05945	-0.156	-0.1599	-0.0351	0.02071	-0.0134	-0.2219	0.0604	-0.1471	-0.2158	-0.0508	0.50877	1

V. EXPERIMENTAL RESULT

This paper employs the following predictive models on the OLAUD dataset mentioned previously to anticipate dropouts. An 80:20 train-test split was utilized to divide the dataset for training and testing. By evaluating the models' performance through different metrics, including accuracy, precision, recall, and F1-score, we aimed to determine the best-suited model for predicting dropouts.

Decision Tree: A decision tree is a type of predictive model that is used in supervised learning to make predictions by recursively partitioning the input space into regions based on the value of different features.

Random Forest: Random forest is an ensemble method that combines multiple decision trees to improve the accuracy and robustness of the model.

Gaussian Naive Bayes: Gaussian Naive Bayes is a probabilistic model that is based on Bayes' theorem and assumes that the features are independent and normally distributed.

AdaBoost Classifier: AdaBoost is an ensemble method that combines multiple weak learners to create a strong learner. It works by iteratively adjusting the weights of misclassified samples.

Extra Tree Classifier: Extra Trees is an ensemble method that is similar to Random Forest, but it uses a different method to generate random splits at each node.

XGBoost Classifier: XGBoost is an ensemble method that uses gradient boosting to improve the accuracy of the model. It is known for its high performance and has won several machine learning competitions.

Multilayer Perceptron (MLP): MLP is a type of feedforward artificial neural network that is used for supervised learning. It consists of multiple layers of nodes and can learn complex non-linear relationships between the input and output variables.

The objective of the task was twofold: to predict whether a student would pass or fail the course and to

determine whether they would drop out or continue with the course. At first, all the predictive models were trained on the training data to forecast whether a student would be PASS or FAIL in the course. The PASS and distinction results were grouped together to form a PASS class, while the FAIL and withdrawn results were grouped together to form a FAIL class. Subsequently, all of the previously mentioned predictive models were trained on the training data to predict whether a student would DROPOUT or NON-DROPOUT from the course. The PASS, FAIL, and DISTINCTION results were grouped together to create a non-dropout class, whereas withdrawn was categorized as a dropout class. Finally, the models were put to the test using the test data to anticipate the students' performance.

Table 1 displays the results of the predictive models after being evaluated with demographic data, clickstream data, and assessment scores. The objective was to predict the final result using the remaining variables as input. The table provides information on precision, recall, accuracy, and f1-score values of the models across different categories of students' final results using the data mentioned above.

TABLE 2:
PASS/FAIL PREDICTION

	Precision		Recall		F1-score		Accuracy
	PASS	FAIL	PASS	FAIL	PASS	FAIL	
Decision Tree classifier	0.82	1.00	1.00	0.31	0.9	0.48	84%
Gaussian naive bayes	0.87	0.61	0.89	0.57	0.88	0.59	81%
Random Forest classifier	0.76	0.00	1.00	0.00	0.86	0.00	76%
Extra tree classifier	0.82	1.00	1.00	0.30	0.90	0.46	83%
XGB classifier	0.89	0.91	0.98	0.62	0.93	0.74	89%
Adaboost classifier	0.86	0.80	0.96	0.50	0.91	0.61	85%
MLP	0.87	0.84	0.97	0.54	0.92	0.65	86%

Table 2 displays the results of the predictive models that were applied to the demographics data, clickstream data, and assessment scores. The goal of the models was to predict the dropout variable, while all other variables served as input. The table shows the precision, recall, accuracy, and f1-score values of the predictive models, which indicate the models' ability to determine if a student will drop out or not based on demographics data, clickstream data and assessment score

TABLE 3:
DROPOUT/NON DROPOUT PREDICTION

	Precision		Recall		F1-score		Accuracy
	Non-Dropout	Dropout	Non-Dropout	Dropout	Non-Dropout	Dropout	
Decision Tree classifier	0.93	0.00	1.00	0.00	0.96	0.00	91.50%
Gaussian naive bayes	0.97	0.18	0.73	0.75	0.83	0.29	73.23%
Random Forest classifier	0.93	0.00	1.00	0.00	0.96	0.00	92.50%
Extra tree classifier	0.93	0.00	1.00	0.00	0.96	0.00	90.50%
XGB classifier	0.95	0.82	0.99	0.33	0.97	0.47	94.43%
Adaboost classifier	0.93	0.51	1.00	0.03	0.96	0.06	92.51%
MLP	0.93	0.67	1.00	0.02	0.96	0.04	92.58%

As per the results displayed in Table 1 and Table 2, the XGBoost classifier performs best with an accuracy of 89% in forecasting whether a student will pass or fail. On the other hand, it has the highest accuracy of 94% when predicting whether a student will drop out or continue the course.

VI. CONCLUSION AND FUTURE SCOPE

While MOOCs have become increasingly popular over the past decade they are facing high dropout and failure rates. The Open University Learning Analytics Dataset (OULAD) is used to evaluate the students' performance using learning behavior pattern in MOOCs with the help of machine learning technique. The dataset contains data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules). In this paper. the OULAD dataset is used to train and test Decision Tree, Random Forest, Gaussian naive bayes, AdaBoost classifier, Extra tree classifier, XGBoost classifier, Multilayer Perceptron (MLP is a feedforward artificial neural network) models techniques to predict student dropout, as well as if students that do not drop out will pass or fail the course.

From the numerical results, we can make the following empirical observations:

1. The XGBoost classifier performs better than others with an accuracy of 89% in forecasting whether a student will pass or fail.
2. At the same time, the XGBoost has the highest accuracy of 94% when predicting whether a student will drop out or continue the course.

Result gives higher accuracy for both models, but there is vast difference between f1-score of positive class and negative class. This difference suggests that the results may be skewed towards the majority class. This phenomenon is known as class imbalance, where one class is much more prevalent in the dataset than the other. So future research may include applying some balancing technique to data before performing training of model.

REFERENCES

1. *Lemay, D.J. and T. Doleck, Predicting completion of massive open online course (MOOC) assignments from video viewing behavior. Interactive Learning Environments, 2020: p. 1-12.*
2. *Feng, W., J. Tang, and T.X. Liu, Understanding Dropouts in MOOCs. Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 33(01): p. 517-524.*
3. *Mubarak, A.A., H. Cao, and S.A.M. Ahmed, Predictive learning analytics using deep learning model in MOOCs' courses videos. Education and Information Technologies, 2021. 26(1): p. 371-392.*
4. *Sanchez-Gordon, S. and S. Lujan-Mora, How Could MOOCs Become Accessible? The Case of edX and the Future of Inclusive Online Learning. Journal of Universal Computer Science, 22(1), 55–81., 2016. 22(1): p. 55-81.*
5. *Laura, P., The year of the MOOC. The New York Times, in The New York Time 2012.*
6. *Gupta, R. and N. Sambyal, An understanding Approach towards MOOCs. International Journal of Emerging Technology and Advanced Engineering, 2013. 3(6): p. 312-315.*
7. *Mourdi, Y., et al., A Machine Learning Based Approach to Enhance MOOC Users' Classification. Turkish Online Journal of Distance Education, 2020. 21(2): p. 47-68.*
8. *Onah, D.F., J. Sinclair, and Boyatt. Dropout rates of massive open online courses: Behavioral pattens of MOOC droput and completion: Existing evalutaion. in 6th International Conference on Education and New Learning Technologies (EDULEARN14). 2014.*
9. *Gardner, J. and C. Brooks, Student success prediction in MOOCs. User Modeling and User-Adapted Interaction, 2018. 28(2): p. 127-203.*
10. *Boyer, S., et al. Data science foundry for MOOCs. in 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2015.*
11. *Taylor, C., K. Veeramachaneni, and U.M. O'Reilly, Likely to stop? Predicting stopout in massive open online courses. arXiv:1408.3382 [cs.CY], 2014.*
12. *He, J., et al., Identifying At-Risk Students in Massive Open Online Courses. Proceedings of the AAAI Conference on Artificial Intelligence, 2015. 29(1).*
13. *Kizilcec, R.F., C. Piech, and E. Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. in Proceedings of the third international conference on learning analytics and knowledge. 2013.*
14. *Mubarak, A., H. Cao, and W. Zhang, Prediction of students' early dropout based on their interaction logs in online learning environment. Interactive Learning Environments, 2020: p. 1-20.*
15. *Costa, E.B., et al., Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, 2017. 73: p. 247-256.*
16. *Baker, R., et al., Does Inducing Students to Schedule Lecture Watching in Online Classes Improve Their Academic Performance? An Experimental Analysis of a Time Management Intervention. Research in Higher Education, 2019. 60(4): p. 521-552.*