RESEARCH ARTICLE OPEN ACCESS

Dropout Prediction in MOOCs Dataset Using Decision Tree and XGB Classifier Algorithm

Kinjal K Patel[1]
1 Lecturer, Government Polytechnic, Ahmedabad
Email: cekinjalvp@gmail.com

Abstract:

MOOCs have become a significant trend in higher education, providing open access to quality educational resources for learners around the world. Their appeal lies in the ability to reach a large audience regardless of geographical location or financial resources. These courses typically feature video lectures, problem sets, discussion forums, and other online tools, offering flexibility and an affordable way to gain new knowledge or skills. Indeed, the growth of MOOCs has been remarkable, with millions of users worldwide. However, the high dropout rate is a significant and persistent challenge. Typically, many learners' sign up for courses, but a large percentage fail to complete them. This phenomenon has led researchers and educators to focus on dropout prediction and strategies for addressing it. This paper focuses on predicting student dropout using machine learning algorithms like decision trees and the XGBoost (XGB) classifier, which are well-suited for this task due to their ability to handle complex patterns and non-linear relationships in data. The use of the OULAD dataset (Open University Learning Analytics Dataset) is particularly valuable, as it provides rich, real-world data on student engagement in online courses.

Key words: Dropout Prediction, MOOC, Learning Analytics, Machine Learning, OULAD Dataset..

INTRODUCTION

We are living in the digital era, where technology plays a central role in how education is accessed and delivered. As the world becomes more dynamic and interconnected, students increasingly turn to digital platforms like Massive Open Online Courses (MOOCs) to acquire new skills and knowledge [1]. MOOCs have become popular around the world in recent years. MOOCs exist on the Internet, which are free from the limitations of traditional classroom teaching time and place, and have more flexible requirements on the learning environment [2] . These kinds of online platforms are acting as ideal choices because of their nature such as flexibility and accessibility, lifelong learning, cost-effective education, global reach and diversity, interactive and engaging learning from diverse educationalists through elite universities [3]. Various famous MOOC providers like HarvardX, Coursera, Unacademy, and NPTEL are extensively used by the students and the working group of all ages to gain knowledge to withstand the competitive world [2, 4, 5].

In India MOOC has been developed rapidly with the recent achievement of launching Swayam [5]. Swayam platform is different from the other MOOCs, it provides credit based on the UGC 2016 framework and students can take the course in Swayam. To earn the certificate, the student has to approach a local institute for the exam and earn the certificate, so it combines both the traditional education system as well as the MOOC concept of education, and the strength of Swayam is that it uses a Qualitative approach to evaluate a student's grade [5].

While MOOCs have gained immense popularity and provide accessible education to millions globally, they still face several significant challenges. One of the most pressing issues is the high dropout rate, which contrasts starkly with traditional face-to-face education. Here are some of the core challenges that contribute to this problem: Lack of Face-to-Face Interaction, Self-Discipline and Motivation, Overwhelming Course Load, Lack of Peer Interaction. [6]. Dropout rates in MOOCs

are generally higher compared with face-to-face education [7]. Student dropout is one of the main problems in MOOCs that has received considerable attention from the scholars across the world [6, 8]. Potential solutions to address these issues to reduce dropout rates, several interventions and strategies can be employed. Some strategies like personalized learning paths, increased interaction, gamification and rewards, early intervention etc[9]. Early detection of students at risk of dropout lays an essential role in reducing the problem, enabling targeted actions aimed at specific situations [10]. Therefore, accurate prediction of whether students drop out of course is helpful for teachers to provide timely intervention, improve the learner's learning effect and realize the platform's [8].

Researchers and technologists have been exploring various Machine Learning (ML) techniques to predict and mitigate student dropout rates in MOOCs. By analyzing patterns in student behavior, interaction data, and demographic information, ML models can identify at-risk students and enable timely interventions to keep them engaged [11]. These techniques have statistically proved to obtain high dropout prediction accuracy.In this paper, decision tree and XGB classifier technique is used for dropout prediction on the OULAD dataset published by Open University of UK.

Methodology

This section, presents the methodology used to carry out this research work and the different tools that contributed to the development of the framework.

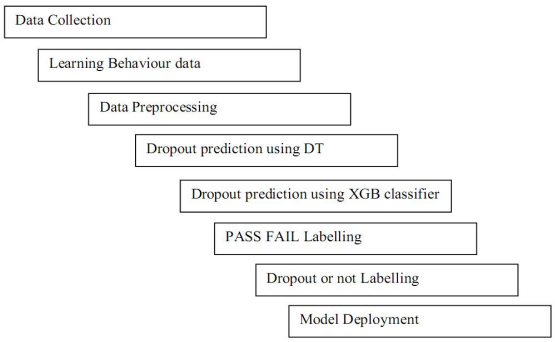

Figure 1: Overall framework of dropout prediction

The proposed approach for predicting student dropout using machine learning techniques is methodically structured, starting with a comprehensive data pre-processing phase. This phase is crucial because it ensures the quality and consistency of the data before feeding it into machine learning models. The first phase is data pre-processing, in which the data extraction, reconstruction, normalization, standardization, scaling, data encoding and necessary transformations of the data are carried out taking into account the characteristics set by the experts. After extraction of data, missing data is handled using different techniques. The data are standardized to unify the units of measurement and thus make possible the comparison between the data. After pre-processing, the data is ready for the model generation phase, where machine learning algorithms like decision trees, XGBoost, or SVM can be trained on the pre-processed data. The clean and well-structured data ensures that the model can focus on finding patterns and relationships without being mislead by noise or inconsistencies in the data[12]. The quality of the pre-processing phase is essential to ensure accurate, efficient, and interpretable machine learning models, leading to more effective dropout prediction and early intervention strategies.

A decision tree classification model is a widely used and powerful machine learning algorithm, especially for classification tasks [13]. A decision tree is a non-parametric model, meaning it doesn't assume any specific distribution for the data, which makes it flexible for various types of data. The goal is to build a model that learns simple decision rules from the data's features in order to predict the target variable, which in this case is whether a student will drop out or continue in the MOOC [14]. The decision tree algorithm works by recursively partitioning the dataset based on feature values until it reaches the leaf nodes. An individual component of the feature vector is looked at in each branching node of the graph. The left branch is followed if the feature's value falls below a predetermined threshold; otherwise, the right branch is followed. The choice of the class to which the example belongs is made as the leaf node is reached [13-15] .

Multi-class classification can be done using decision tree classifiers on a dataset. As with other classifiers, it takes as input two arrays: an array X, sparse or dense, of shape (n_samples, n_features) holding the training samples, and an array Y of integer values, shape (n_samples,), holding the class labels for the training samples [14]. The model can then be used to forecast the class of samples after being fitted [13].

Binary classification using labels of [-1, 1] and multiclass classification using labels of [0,..., K-1] are both possible using decision tree classifiers.

The XGBoost (Extreme Gradient Boosting) classifier is a powerful and efficient machine learning algorithm widely used for classification tasks, including predicting student dropout in MOOCs[16]. It belongs to the ensemble learning family of algorithms and is known for its excellent performance and scalability, particularly on large and complex datasets. So XGB classifier algorithm is trained and tested on the same dataset. By capturing complex relationships between features, XGBoost can help identify at-risk students early and support strategies to improve retention.

Dropout prediction is conducted using data from the Open University Learning Analytics Dataset (OULAD). OULAD is a widely used dataset for educational data mining and learning analytics research, containing comprehensive information about student interactions, assessments, and demographic details within the Open University's online courses [17]. OULAD dataset was published by the Open University, UK. A public British university, Open University has the most undergraduate students in the UK. As implied by the name, the majority of students at Open University attend classes off-campus.

Data from 32,592 students and 22 module presentations are included in the OULAD dataset. It includes information on seven chosen courses' courses, students, and their interactions with virtual learning environments (VLEs) (called modules). Course presentations begin in February and October, and are denoted by the letters "B" and "J," respectively. The collection consists of connected tables with distinctive identifiers. The CSV format is used to store all tables. The data that are utilised for analysis are described in the section that follows.

The OULAD dataset includes various types of data related to students' learning behaviors, which are crucial for predicting dropout. It covers 1)Demographic Information which Includes features such as student age, gender, region, and prior educational background, which provide insights into how student characteristics might influence dropout rates. 2) Assessment Information like Data on student performance in formative and summative assessments (e.g., scores, completion status) across different courses, it also provides an indication of how well a student is engaging with the course material and performing academically. 3) Interaction with Virtual Learning Environment (VLE) which Tracks student interactions with the online platform, such as logins, page views, and resource usage and this type of data gives a detailed view of a student's activity and engagement patterns, which are critical factors in dropout prediction. 7 different csv files Courses, Assessments, Vle, studentInfo, studentRegistration, studentAssessment, studentVle are used for data modelling.

Data Preprocessing of OLAUD dataset

The process of transforming raw data into a usable dataset is known as preprocessing, and it is a critical step in any machine learning project[18]. Preprocessing is often a labor-intensive task that requires a combination of creativity**,** technical skills, and, ideally, domain knowledge from the data analyst. The primary goal of preprocessing is to create a set of informative features that will allow the learning algorithm to build an accurate and effective predictive model [19]. Many different methods like One-Hot Encoding, Binning, Normalization, Standardization, Dealing with Missing Features, Data Imputation Techniques can be used for data pre-processing [14].

Here, data preprocessing involves handling the missing values in the Open University Learning Analytics Dataset (OULAD)**,** which contains seven CSV files that store various types of student information and interaction data. To ensure the quality and consistency of the dataset, different techniques are applied to handle missing values based on the context of each file and the proportion of missing data. In files where a small percentage of

rows contain missing values compared to the total dataset, these rows are simply removed. For instance, in the studentAssessment file, where the number of rows with missing data is negligible, those rows are dropped to maintain data integrity without significantly affecting the overall dataset. In cases where missing values are more substantial but still manageable, they are filled with appropriate statistical measures like the mean (for numerical data) or mode (for categorical data). In the studentRegistration file, some missing values in the data_registration column (which tracks the registration date) are replaced with 0. This allows for the retention of important information while filling in gaps without losing rows. In the studentInfo file, the imd_band column (which represents categorical data indicating the student's socio-economic background) contains missing values. Since this is a categorical feature, the missing values are filled using the mode (the most frequent category). This ensures that the data remains consistent while minimizing the impact of missing values.

Vle file contains different activity types like resource, subpage, oucontent, url etc. So count plot is created for different activity types. It can be seen from the graph that 'resource' has the most data points in vle file.

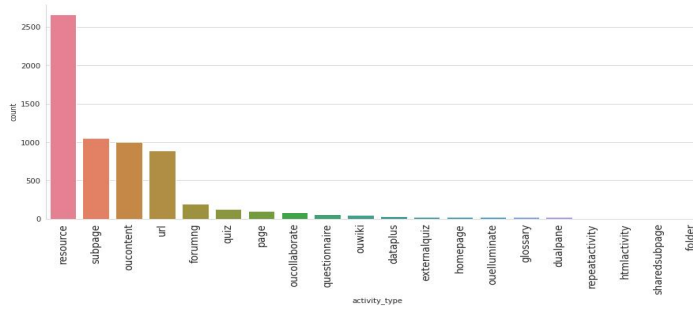

Figure 2: Count plot of different activity types in vle data

By considering day of click three features are created click_timing, before_click and after_click. There would have been instances where students submitted their assignments later than the deadline. Based on the columns date_submitted and date, one column is created of late submission or not for the student.

There would have been instances where students su bmitted their assignments later than the deadline. So late submission or not is created using the date_submitted (days after student submitted their assignment) column and date (deadline in days for the assignment). Here 0 means Late submission and 1 means On time. Then different data files are merges as per the requirements. There are two types of categorical variables in the data nominal (here there is no order in the category) and ordinal (here there is order in category). Categorical encoding is performed on these data. Now, data are ready for model.

One count plot is generated for counting number of students 'PASS', 'FAIL', 'WITHDRAWN', 'DISTINCTION' as shown in Chart 2.

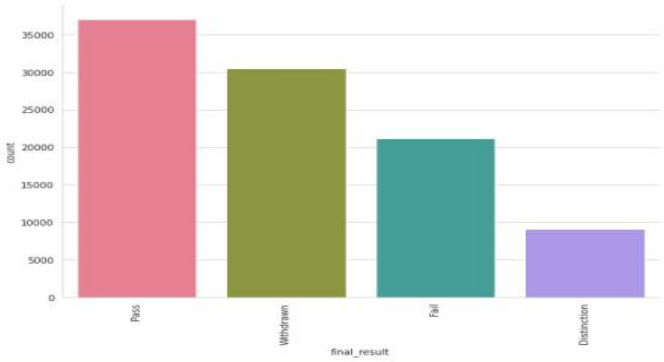

Figure 3 : Count plot for students result

In this paper, fail and withdrawn are considered as fail. Similarly, pass and distinction are considred as pass in final_result. A dropout column is created which shows 1 for dropout and 0 for not dropout.

Dropout prediction using Decision Tree algorithm

There are two definitions of MOOC dropout prediction in the current studies. The first is whether a learner will "PASS" or "FAIL" in the course and the second is learner will "DROPOUT" or "NOT DROPOUT" in the course. Three different types of information are used for creation of model. The first one is student demographic data which

gives student general information like gender, region, age, highest education, no of previous attempts etc. The second one is student interaction with Virtual Learning Environment(VLE) , that is no of times student has interacted with material and other resources. The third one is assessment information, that is student score in each and every assessment. The decision tree implemented and the feature selection is performed using decision tree based on Gini Index. Decision tree is constructed by recursive partitioning into smaller subsets until reaching the specified stopping criterion, for example, that all the subsets belong to a single class. A single feature split is recursively defined for all nodes of the tree using some criterion[18]. Gini Index is one of the most widely used criteria for decision tree. Gini index is a measure of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm. An attribute with the low Gini index should be preferred as compared to the high Gini index. It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits. Gini index can be calculated using the below formula:

$$Gini(t) = 1 - \sum_{i=1}^{j} P(i|t)^2$$

Where,
The j represents the number of classes in the label, and
The P represents the ratio of class at the ith node.
Gini impurity has a maximum value of 0.5, which is the worst we can get, and a minimum value of 0 means the best we can get. Based on this Gini impurity root node, intermediate node and leaf nodes are finalizes and decision tree is created. This fitted decision tree is used to test data and to find out whether student will "PASS or FAIL" and "DROPOUT or NOT".

Decision Tree formation and Analysis

The whole dataset is divided into 80-20 ratio for training and testing of model. As explained in previous topic decision tree classification algorithm is trained on 80% dataset and the following tree for student will "PASS" or "FAIL" is created.

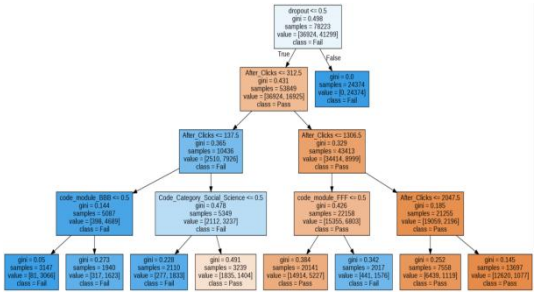

Figure 4: Decision Tree for class Label "PASS" or "FAIL"

Similarly, another tree for student will "DROPOUT" or "NOT" is created.

Dropout prediction using XGB classifier algorithm

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels. There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees).
XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library[18]. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

Evaluating Model

Two machine learning techniques Decision tree and XGB classifier are used as predictive model that analyses the problems faced by at-risk students, subsequently, facilitating instructors for timely intervention to persuade students to increase their study engagements and improve their study performance. The predictive model is trained and tested using machine learning (ML) algorithms to characterize the learning behaviour of students according to their study variables. The performance of both ML algorithms is compared by using accuracy, precision, support, and f-score. The predictive model can help instructors in identifying at-risk students early in the course for timely intervention thus avoiding student dropouts. Our

results showed that students' assessment scores, engagement intensity i.e. clickstream data, and time-dependent variables are important factors in online learning.

All demographics variables, assessment variables and click stream data were considered for training Decision Tree and XGBoost classifier algorithams. First of all above predictive models are trained using train data then PASS or FAIL is predicted using same trained models. Here pass and distinction result are combined into PASS class whereas fail and withdrawn are combine into FAIL class. After that all above predictive models are trained using train data and then DROPOUT or NON DROPOUT is predicted using same trained models. Here pass, fail and distinction result are combined into non dropout class whereas withdrawn is considered as dropout class.

Table 1: PASS/FAIL prediction

	Precision		Recall		F1-score		Acc
	PASS	FAIL	PASS	FAIL	PASS	FAIL	
DT	0.82	1.00	1.00	0.31	0.9	0.48	84%
XGB	0.89	0.91	0.98	0.62	0.93	0.74	89%

Table 2: Dropout/Non dropout prediction

	Precision		Recall		F1-score		Acc
	continue	dropout	continue	dropout	continue	dropout	
DT	0.93	0.00	1.00	0.00	0.96	0.00	93%
XGB	0.95	0.82	0.99	0.33	0.97	0.47	94%

Discussion

Here data are divided into training and testing dataset. I have keep 80% of the data for training and only 20% for testing. All demographics data, clickstream data and assessment data are used to train and test decision tree and XGB classifier models. Classification of students is predicted in two phases, first student will pass or fail and in second phase it predicts that student will dropout or continue. Here the decision tree algorithm gives an accuracy of 84% for Pass/Fail prediction while XGB classifier gives an accuracy of 89% for Pass/Fail prediction. The decision tree algorithm gives an accuracy of 93% for dropout or non dropout prediction while XGB classifier gives an accuracy of 94% for dropout or non dropout prediction.

Conclusion

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). The OULAD dataset is analysed using decision tree classification and XGB classifier for pass/fail and dropout/non dropout binary output. Result shows that the XGB classifier is more accurate than the decision tree classifier. As XGB classifier improves the result as it is an ensemble learning technique in which trees are built in parallel.

References

1.Barak, M., A. Watted, and H. Haick, *Motivation to learn in massive open online courses: Examining aspects of language and social engagement.* Computers & Education, 2016. **94**: p. 49-60.

2.Gupta, R. and N. Sambyal, An understanding Approach towards MOOCs. International Journal of Emerging Technology and Advanced Engineering, 2013. 3(6): p. 312-315.

3.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.

4.Dalipi, F., A.S. Imran, and Z. Kastrati, MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges, in Educaon. 2018.

5.Kaveri, A., et al., *Decoding the Indian MOOC learner*, in *IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*. 2015: Amritsar. p. 182-187.

6.Wan, H., et al. *Dropout Prediction in MOOCs using Learners' Study Habits Features*. in *EDM*. 2017.

7.Xing, W., et al., *Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization.* Computers in Human Behavior, 2016. **58**: p. 119-129.
8.Xing, W. and D. Du, *Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention.* Journal of Educational Computing Research, 2018. **57**(3): p. 547-570.
9.Ren, Z., H. Rangwala, and A. Johri, *Predicting performance on MOOC assessments using multi-regression models.* CoRR, vol. abs/1605.02269, [Online]. Available: http://arxiv.org/abs/1605.02269] 2016.
10.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.
11.Al-Shabandar, R., et al. *Machine learning approaches to predict learning outcomes in Massive open online courses*. in *2017 International Joint Conference on Neural Networks (IJCNN)*. 2017.
12.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.
13.Trappenberg, T., *Fundamentals of Machine Learning*. 2019: OUP Oxford.
14.Mitchell, T.M., *Machine Learning*. 1997: McGraw-Hill.
15.Chen, J., et al., *MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine.* Mathematical Problems in Engineering, 2019. **2019**: p. 8404653.
16.Nagrecha, S., Dillon, J.Z., Chawla, N.V. *MOOC Dropout Prediction: Lessons Learned from Making Pipelines Interpretable*. in *26th International Conference on World Wide Web Companion*. 2017. Republic and Canton of Geneva, Switzerland.
17.Kuzilek, J., M. Hlosta, and Z. Zdráhal, *Open University Learning Analytics dataset.* Scientific Data, 2017. **4**: p. 170171.
18.Gardner, J. and C. Brooks, *Student success prediction in MOOCs.* User Modeling and User-Adapted Interaction, 2018. **28**(2): p. 127-203.
19.Burkov, A., *The Hundred-Page Machine Learning Book*. 2019: Andriy Burkov.