Feature Engineering for Predicting Student Dropout in Massive Open Online Courses

Lopa Mandal[1*], and Aneesh Kar[2]

[1] *Alliance University, Department of Computer Science and Engineering, Bengaluru, India*

[2] *Institute of Engineering & Management, Department of Infomation Technology, Kolkata, India*

Abstract: Massive open online courses (MOOCs) were invented to impart education to people who could not afford or get access to traditional education options. In recent times there is an unprecedented shift to online platforms for teaching-learning around the world. Dropout prediction or identifying students at risk of dropping out of a course, therefore becomes an important problem to study due to the high attrition rate commonly found on many MOOC platforms. Proper feature engineering plays an important role in getting desired results in such predictions. The present work thus focuses on providing valuable features for improving the prediction results. These predictions are then performed and analyzed by Machine Learning algorithms. The Random Forest Classifier gave 86.14% accuracy and outperformed Logistic Regression, Decision Tree Classifier, Gradient Boosting Classifier and Xgb Classifier applied in the present work.

Keywords: MOOCs, feature engineering, data analytics, Dropout Prediction, Machine Learning.

1. Introduction

Massive open online courses (MOOCs) have revolutionized education by providing a platform for anyone to access courses offered by top universities and institutions from anywhere in the world. Millions of people around the world use MOOCs to learn for a variety of reasons, including career development, changing careers, college preparations, supplemental learning, lifelong learning, corporate eLearning & training, and more. Massive open online courses, or MOOCs, have seen a sharp rise in the number of enrolments ever since March 2020., when most of the countries began the COVID-19 enforced lockdown. Coursera – a popular online platform that offers MOOCs, has skyrocketed and was 640% higher from mid-March to mid-April than during the same period last year, growing from 1.6 to 10.3 million. The surge was driven in part by giving free catalog access for 3,800 courses to their university partners. MOOCs allows self-paced learning and as a result it shows a much wider range of engagement patterns. Unlike the traditional brick-and mortar classrooms, students are not always directly in touch with human teacher and as a result many of them are not motivated enough to complete a course. This leads to the undeniable fact that MOOCs do observe an extremely high dropout rate.

Existing works in the related field have been studied to identify different factors affecting drop-out rates in MOOCs platforms, techniques used for drop-out prediction and the various challenges faced in predicting dropouts.

It has been found out that student demographics such as age, gender, and educational background can affect drop-out rates [1, 2]. Course content difficulty, course workload, and course relevance to students' interests are also identified as significant factors in predicting dropouts. Student engagement, measured from clickstream data by the number of videos watched, the number of forum posts, and the number of quizzes attempted over several consecutive days has also been found to be significant predictors of dropouts by many researchers [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. Past and current activities in addition to past performance, behavioral features, followed by temporal and demographic features, influence of friend/colleagues are also significant factors [2, 12, 7].

Though there are progresses found in drop-out predictions in MOOCs platform, many challenges still remain. There are difficulties in collecting sufficient data on student interaction with the MOOCs platform. Moreover, data collection method is different in different platforms, making it difficult to compare results across MOOCs platforms.

Many researchers mentioned that there is no proper accepted definition of dropout for MOOCs and different definitions can be adapted. Most commonly used predictors of dropout include time spent on the course, interaction pattern with the MOOCs platform, demographic factors etc. It has also been found out that Machine learning approaches outperform traditional statistical models for the prediction of student dropout in MOOCs and the overall effectiveness of a dropout prediction model varies depending on various characteristics of the MOOC. The studies revealed that personalized interaction with the learners can improve retention rates in MOOCs.

Considering the above-mentioned points, the motivation behind the present work is set to predict the dropout of students beforehand so that the course providers can take certain measures to prevent them from dropping out of their respective courses. Proper feature engineering is extremely important in getting a desired result and this paper focuses on providing valuable features for improving the prediction results.

2. Materials and methods

2.1 Background study for techniques used for drop-out prediction

Researchers have used various techniques for predicting dropouts in MOOCs. Machine learning and data mining techniques have been found to be popular in predicting dropouts. Some researchers used a decision tree model to predict student drop-out in a programming MOOC. The study considered that if learners do not have any learning activities in ten continuous days, there is a high chance of being drop-out. They used multiple supervised classification models such as SVM, logistic regression, random forests, and gradient boosting decision trees. They thereby achieved 88% accuracy in drop out prediction task with GBDT model [5, 14, 2]. Process mining and sequence mining techniques using log of interactions of learners with the courses in the MOOC platform also gave good predictions. Statistical analysis shows that there is high correlation between dropouts of different courses. It also revealed that friends' dropout behaviors play significant role. With these insights, a group of researchers proposed a Context-aware Feature Interaction Network (CFIN) for predicting learner's dropout behavior [7]. Another research focused on an approach to early predicting which students are at risk of missing assignments in online courses. They used neural networks and considered activity records of students with the learning management systems (LMS), assessment methods of different activities, course structures etc. to be important factors affecting student drop-out [15, 5].

A group of researchers worked on the datasets from the MOOCs of Peking University running on the Coursera platform, from which they extracted 19 major features of tune in after analyzing the log structure including the characteristics of learners based on learners' start and dropout time statistics. They used various machine learning approaches to create a sliding window model for predicting the probability of abandonment on a given course. They claimed that the machine learning and sliding window model could actually predict the loss of students with a high accuracy [6].

A study on temporal models for predicting student dropout in MOOCS mentioned that there is no proper accepted definition of dropout for MOOCs and different definitions can be adapted. The study came up with three definitions –participation in the final week, last week engagement and participation in the next week by tracking learners' access to lecture videos and their forum activities on MOOC platform during the period of their course. RNN model with LSTM cells was used and significantly better results were obtained over the other methods used on the definitions adapted in this work [16].

A study in 2018 proposed an end-to-end dropout prediction model based on two-dimensional convolutional neural networks (DP-CNN) and use this model to directly analyze clickstream data to make a final prediction. The authors concluded that the DP-CNN model can directly process the clickstream data and automatically extract features for the final prediction with reduced the complexity of feature extraction and improved feature quality. They used the dataset of 39 courses collected from XuetangX, a dataset on which much previous research work have been performed [17]

Josh Gardner, Yuming Yang, Ryan S. Baker and Christopher Brooks, in 2019 conducted a three-part replication case study of a state-of-the-art LSTM dropout prediction model. They demonstrated that their process could reduce overfitting and improve generalization performance of the model but could not achieve the previously-reported level of performance. Their work proposed a paradigm of end-to-end reproducibility using the MOOC Replication Framework [18].

Jacob Whitehill Kiran Mohan, Daniel Seaton, Yigal Rosen and Dustin Tingley, in 2017, compared the accuracy of mainstream dropout prediction architectures primarily under four different training paradigms. Their results suggested that training and testing on the same course can overestimate accuracy by several percentage points. Their classifier performance didn't vary significantly across disciplines. They also investigated a novel dropout prediction architecture based on fully connected deep feedforward neural networks and found that networks with up to five hidden layers improved test accuracy over logistic. One regression an have a statistically significant increase [19].

Another work explored ensemble, deep learning, and regression techniques for predicting dropout and outcome in MOOCs using the Open University Learning Analytics Dataset (OULAD). This is a large and complex dataset contains various data such as student demographic information, assessment data, and more. In this work the researchers used the Knowledge Discovery in Databases (KDD) methodology. They built different models based on different categorical attributes e.g. demographic information, reputation scores, interaction data with Virtual Learning Environment (VLE) etc. The results show that the machine learning model based on the students' interaction with their VLE performed well. Their results also showed a slight improvement when considered student demographic information and assessment scores, as well as VLE interaction. They ultimately said their future focus will be on feature selection and engineering, including time-based metrics related to assessment and student interaction, to improve dropout and outcome prediction performance [20]. Some other deep learning-based works also gave promising result [21, 3, 22, 23]. Some researchers proposed to handle dropout prediction by merging global and local tensors to represent aspects of all available features. A global tensor structure was proposed to model the MOOC data, while a local tensor was clustered to represent course connections. A new similarity estimation method was then introduced to enhance the explanatory power of the cluster. The model produced highly satisfactory results [24].

2.2 Proposed methodology

To proceed with the work and for the experimental setup, the collection of data is extremely important. The dataset used in the present work is taken from MOOC platform, XuetangX [25]. The chosen dataset has the rows which are time logs for a particular action undertaken by a particular learner at a given time. Data preprocessing is explained in two sections separately for course data and students' data.

2.2.1 The preprocessing and feature engineering of course data

Table 1 shows the initial features of the chosen dataset.

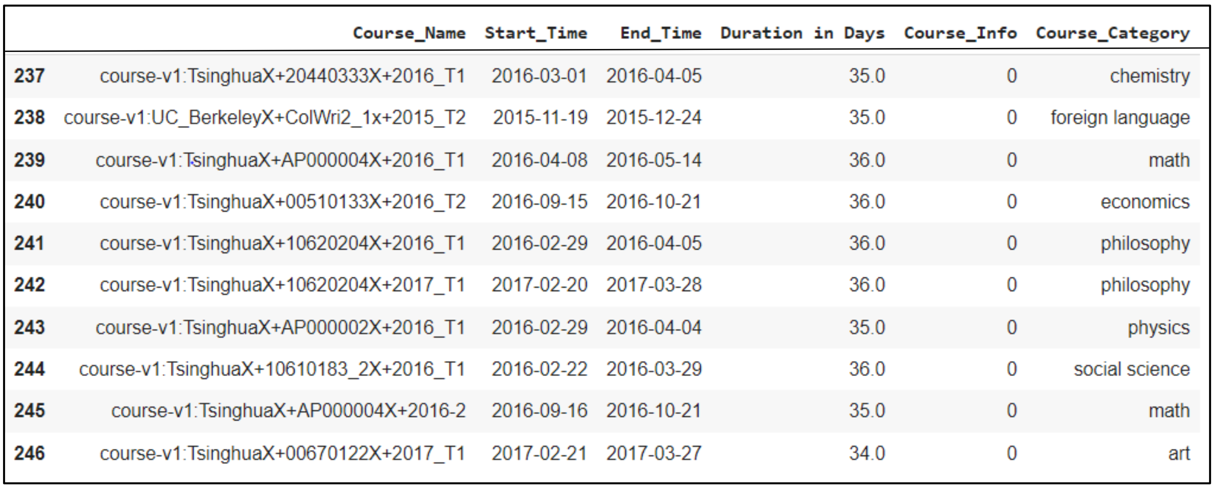
Table 1. Initial features of the dataset of "Course data".

The Original Features	Feature Description
enroll_id	The Unique ID of the (learner & course) pair
username	The Username of the learner
course_id	The Unique ID of the course
session_id	The ID of each session of the learner
action	The type of action the learner is performing at that particular instant.
object	The corresponding Object of the actions of the learner.
time	The occurrence time of the particular action.

Alongside the features mentioned in Table 1, metadata about course information and learners are also available which have later been mapped with the original dataset after extracting meaningful features out of it.

Since the original data came in the form of time logs, performing feature engineering to construct meaningful data/ features, became a challenge. It was given in the source of the dataset taken from MOOC platform XuetangX [25], that the courses considered as instructor paced were tagged as 0. As all the courses which were available for our dataset were tagged with 0, we considered the courses to be instructor paced. In the present work, it has been considered that for an instructor paced course, the last engagement of a learner in that course can be proved to be a major feature in the prediction of his/her dropout. Extracting information regarding the duration of the course was important in finding out the last engagement of the learner. An assumption was made that, since these were instructor paced courses, the difference between the first engagement of any learner for a given course and the last engagement of any learner for the given course will provide a rough estimated time for the duration of the course. To validate this assumption, calculations were carried out and the results hence obtained solidified the assumption. The Course durations were between 32 to 36 days with 36 being most frequent.

The other information related to the courses which included ‘Course_Info’ and ‘Course_Category’ were mapped, imputed and a final dataset about the course information was produced. Refer Figure 1 that displays the attributes of course data after performing preprocessing, feature engineering.



	Course_Name	Start_Time	End_Time	Duration in Days	Course_Info	Course_Category
237	course-v1:TsinghuaX+20440333X+2016_T1	2016-03-01	2016-04-05	35.0	0	chemistry
238	course-v1:UC_BerkeleyX+ColWri2_1x+2015_T2	2015-11-19	2015-12-24	35.0	0	foreign language
239	course-v1:TsinghuaX+AP000004X+2016_T1	2016-04-08	2016-05-14	36.0	0	math
240	course-v1:TsinghuaX+00510133X+2016_T2	2016-09-15	2016-10-21	36.0	0	economics
241	course-v1:TsinghuaX+10620204X+2016_T1	2016-02-29	2016-04-05	36.0	0	philosophy
242	course-v1:TsinghuaX+10620204X+2017_T1	2017-02-20	2017-03-28	36.0	0	philosophy
243	course-v1:TsinghuaX+AP000002X+2016_T1	2016-02-29	2016-04-04	35.0	0	physics
244	course-v1:TsinghuaX+10610183_2X+2016_T1	2016-02-22	2016-03-29	36.0	0	social science
245	course-v1:TsinghuaX+AP000004X+2016-2	2016-09-16	2016-10-21	35.0	0	math
246	course-v1:TsinghuaX+00670122X+2017_T1	2017-02-21	2017-03-27	34.0	0	art

Figure 1. Attributes of “Course data” after preprocessing

2.2.2 The preprocessing & feature engineering of learners’ data

The entire process of extracting and filtering meaningful learner’s data was divided into two major steps. First the preprocessing of learners’ data, followed by the correlation of the learners’ data with the truth values.

2.2.2.1 Preprocessing of learners' data

After the extraction of courses data, especially the course duration (named as ‘CDuration’ in the data set) and the last recorded date for the course in the dataset, the last engagement time of the learners for a given course (‘LastEng’ in the data set) was extracted by finding out the difference between the last recorded date for the course in the dataset and the last recorded engagement date for the particular learner for that course. In the present work, it has been assumed that the ‘LastEng’ feature plays an important role in dropout prediction of a learner. The more the value of the ‘LastEng’ column, more are the chances of that learner to be a dropout.

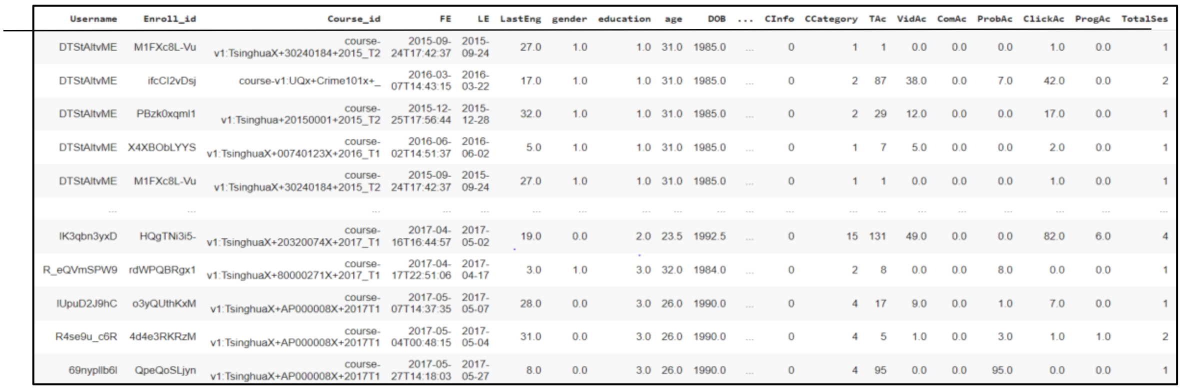
The additional metadata about the learners were also mapped with the original dataset. Information about the age, gender and the educational background of a learner were assumed to give a correlation with truth values of the dataset.

Following the above feature engineering, came another important action which is considered to have a significant contribution to the eventual dropout prediction. The most important feature from the original dataset was the action points of the learner. A learner's actions tell a lot about his/her motivation towards the course. There was a total of 22 possible actions from a learner's perspective, from the dataset. Those are: 'click_about', 'click_info', 'click_courseware', 'load_video', 'close_courseware', 'play_video', 'seek_video', 'pause_video', 'problem_get', ''problem_check', 'problem_check_incorrect', 'problem_check_correct', 'stop_video', 'create_thread', 'create_comment', 'problem_save', 'click_forum', 'click_progress', 'reset_problem', 'delete_thread', 'delete_comment', 'close_forum'. Working with all the 22 actions didn't make any sense hence, these were classified into more meaningful features. Each of the 22 features were further mapped into 4 categories, Video actions, Community Involvement, Problem Access, and Click Action. A separate independent category for Progress Check Action was created. This was an extremely important feature named as 'ProgAC' in the dataset, as more frequently learners check their progress in the course, the chances are that they are interested and is willing to complete it. The other features included are 'TAC', 'vidAc', 'ComAc', 'ProbAc' and 'ClickAc'. 'TAC' was a new feature introduced, which gives us the total actions of the learner for that particular course. More actions, the more likely that the user will finish the course. "ProgAc" includes every time the user clicks on the course's progress bar to check its progress. The present work assumes that a user will only check its progress bar if he/she is interested to finish the course. The more the user checks, the more are the chances to complete the course. Table 2 describes the intention of each of the above-mentioned new features.

Table 2. New Features added to "Learners' data".

Feature	Feature full form	Definition
TAC	Total actions	This gives us the total actions of the learner for that particular course. More actions, the more likely that the user will finish the course.
LastEng	Last Engagement	The difference between the last recorded date for the course in the dataset and the last recorded engagement date for the particular learner for that course.
vidAC	Video Action	This comprises the summation of all the actions related to course videos/ lectures. This included actions like seek video, pause video, play video etc.
ComAC	Community Actions	This includes every action of the learner in community/ forum engagement. This includes actions like the user's involvement in the forums. Posting doubts, etc.
ProbAC	Problems Action	The actions related to the assignments and problems for the particular course.
ClickAC	Click Action	Actions which are related to all the clicking activities on the course's info, about, progress section.
ProgAC	Progress Check Action	This action includes every time the user clicks on the course's progress bar to check its progress. We believe that a user will only check its progress bar if it is interested to finish the course. The more the user checks, the more are its chances to complete the course. This proved to be a major feature in the prediction.

Another important aspect was the total number of sessions the learner undertakes to finish the course. More the number of sessions, more are the chances of the learner to complete the course. The total number of sessions of a learner, named as 'TotalSes' in the dataset, for a particular course was added up.



Username	Enroll_id	Course_id	FE	LE	LastEng	gender	education	age	DOB	...	CInfo	CCategory	TAc	VidAc	ComAc	ProbAc	ClickAc	ProgAc	TotalSes
DTStAltvME	M1FXc8L-Vu	course-v1:TsinghuaX+30240184+2015_T2	2015-09-24T17:42:37	2015-09-24	27.0	1.0	1.0	31.0	1985.0	...	0	1	1	0.0	0.0	0.0	1.0	0.0	1
DTStAltvME	ifcCl2vDsj	course-v1:UQx+Crime101x+_	2016-03-07T14:43:15	2016-03-22	17.0	1.0	1.0	31.0	1985.0	...	0	2	87	38.0	0.0	7.0	42.0	0.0	2
DTStAltvME	PBzk0xqml1	course-v1:Tsinghua+20150001+2015_T2	2015-12-25T17:56:44	2015-12-28	32.0	1.0	1.0	31.0	1985.0	...	0	2	29	12.0	0.0	0.0	17.0	0.0	1
DTStAltvME	X4XBObLYYS	course-v1:TsinghuaX+00740123X+2016_T1	2016-06-02T14:51:37	2016-06-02	5.0	1.0	1.0	31.0	1985.0	...	0	1	7	5.0	0.0	0.0	2.0	0.0	1
DTStAltvME	M1FXc8L-Vu	course-v1:TsinghuaX+30240184+2015_T2	2015-09-24T17:42:37	2015-09-24	27.0	1.0	1.0	31.0	1985.0	...	0	1	1	0.0	0.0	0.0	1.0	0.0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
IK3qbn3yxD	HQgTNi3i5-	course-v1:TsinghuaX+20320074X+2017_T1	2017-04-16T16:44:57	2017-05-02	19.0	0.0	2.0	23.5	1992.5	...	0	15	131	49.0	0.0	0.0	82.0	6.0	4
R_eQVmSPW9	rdWPQBRgx1	course-v1:TsinghuaX+80000271X+2017_T1	2017-04-17T22:51:06	2017-04-17	3.0	1.0	3.0	32.0	1984.0	...	0	2	8	0.0	0.0	8.0	0.0	0.0	1
lUpuD2J9hC	o3yQUthKxM	course-v1:TsinghuaX+AP000008X+2017T1	2017-05-07T14:37:35	2017-05-07	28.0	0.0	3.0	26.0	1990.0	...	0	4	17	9.0	0.0	1.0	7.0	0.0	1
R4se9u_c6R	4d4e3RKRzM	course-v1:TsinghuaX+AP000008X+2017T1	2017-05-04T00:48:15	2017-05-04	31.0	0.0	3.0	26.0	1990.0	...	0	4	5	1.0	0.0	3.0	1.0	1.0	2
69nypllb6l	QpeQoSLjyn	course-v1:TsinghuaX+AP000008X+2017T1	2017-05-27T14:18:03	2017-05-27	8.0	0.0	3.0	26.0	1990.0	...	0	4	95	0.0	0.0	95.0	0.0	0.0	1

Figure 2. Attributes of "Learners' data" after preprocessing

After all the preprocessing and the feature engineering measures were taken, the final dataset comprised 67703 rows of unique enrollments. Figure 2 shows the attributes of the final dataset of learners' data. Please note that the username (username) and the enrollment ID (Enroll_id) have been encoded.

2.2.2.2 Correlation of the learners' data with the truth values

After preprocessing, the effectiveness of the final features for dropout prediction was tested by plotting a heatmap to measure the correlation between the features and their correlation with the truth values of the dataset. The last engagement feature, named as 'LastEng' in the dataset, gave a good correlation with the truth values for the dataset and proved to be a valuable feature. The feature for progress check action named as 'ProgAC' in the dataset also gave a very good correlation with the truth values of the dataset. The features mentioned in Table 2 gave good correlation with the truth values of the dataset and proved to be extremely important for the prediction. The total number of sessions named as 'TotalSes' of a learner for a particular course also proved to be extremely effective in the prediction. Refer Figure 3 for the heap map showing the correlations of all the features with the truth values of the dataset.

3. Experimental setup and results

The dataset used in the present work consisted of the logs of all users' learning activities in a MOOC platform, XuetangX [25]. Python was chosen as the coding language for performing the experiment and Google Colab was used as the environment for conducting the Experiment.

After the completion of the preprocessing, feature engineering and imputation, the final dataset contains 23 properties/ data columns viz. Username, Enroll_id, Course_id, FE, LE, LastEng, gender, education, age, DOB, stime, etime, CDuration, CInfo, CCategory, TAc, VidAc, ComAc, ProbAc, ClickAc, ProgAc, TotalSes, and truth. The final dataset comprised 67703 rows of unique enrollments. The Python Libraries used for plotting the graphs are matplotlib, pyplot, numpy, pandas, and seaborn. A heatmap (refer Figure 3) was plotted to measure the correlation between the features and their correlation with the truth values of the dataset. The heatmap gave a good insight about the correlation of the different features between themselves and most importantly the correlation of the features with the truth column. More the value (negative or positive) for any feature against the truth column, more the weightage it will have in the prediction. The features which did not contribute to the prediction of the model and showed less correlation with the truth column, were dropped. Since the problem statement is a good example of Supervised Classification Machine Learning, the choice of good features played an extremely important role. The features relating to the type of action the learner undertakes, provided good correlation with the Truth values of the dataset. Hence those got more weightage in the prediction. From the heatmap of Figure 3, it is visible that the independent action of progress check action ('Prog_ac' in Figure 3) gave a good correlation with the truth values. Since the courses were instructor paced, the last engagement ('LastEng' in Figure 3) of the students attained more weightage as well.

The preprocessed dataset then applied on different classification models viz. Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier, XGB Classifier, for dropout prediction. Appropriate Python Libraries were used for the purpose.

For each of the Classification Models applied, the dataset was split into 70% Train data, used for training the model and 30% for Test data, used for testing the model, respectively. The Random Forest Classifier gave 86.14% accuracy and outperformed Logistic Regression, Decision Tree Classifier, Gradient Boosting Classifier and XGB Classifier applied in the present work. It showed improvement in results over the state of the art works as well. The results hence received is furnished in Table 3.

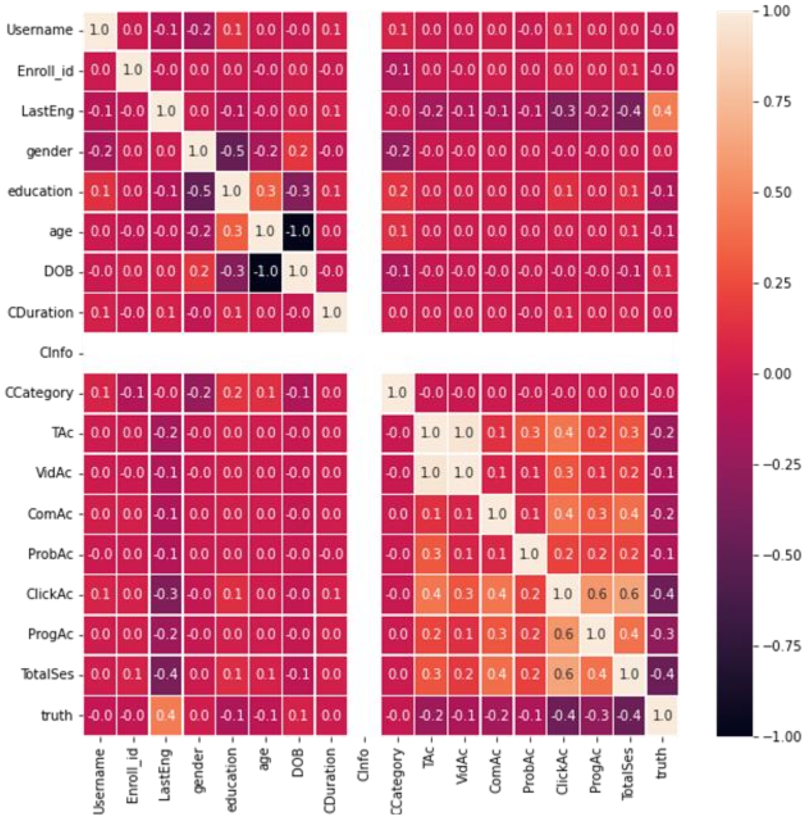


Figure 3. Heatmap measuring the correlation between the features and their correlation with the truth values of the dataset

Table 3. Results of applying the final dataset on different classification models.

Models	Accuracy percentage
Logistic Regression	81.39
Random Forest Classifier	86.14
Decision Tree Classifier	78.32
Gradient Boosting Classifier	85.30
XGB Classifier:	85.76

4. Conclusion and future scope

This present work developed a model to predict dropouts in Massive Open Online Courses (MOOCs). It can be observed that there is a significant dropout in different courses and to prevent that, this work predicts the dropout beforehand.

From the given experimental results, it is observed that there is a significant improvement in percentages, in comparison to the previous works. Especially the Random Forest Classifier, which gave 86.1366974053668% accuracy, which gives a better percentage improvement in comparison to the previous works. After which the XGB Classifier also gave a better performance in comparison to the previous works.

Feature Engineering is extremely important especially when the problem is an example of a Supervised learning classification problem. Choice of correct features helped in achieving a better accuracy in this work. Considering the recent engagement of the students towards the last 7 to 10 days of the course proved to be a major factor in the prediction. Alongside this, the keeping a count of the user actions, especially the user's constant checking of his/her progress proved to be a good feature for the prediction. Based on these historical data, the Course providers can take up measures to prevent their learners from dropping out of the courses.

For future development on this, consideration of peer performance can be used to enhance the predictions. If a group of friends enroll for a particular course, the chances of any one of the members finishing the course can in some way be mapped with the tendency of its immediate peers. Furthermore, historical information regarding the performance of the course on the course providing platform can help in better prediction. Another aspect which can prove to be effective is having the historical information of the users. If a learner has shown interest and completed a course in a category, then the chances of completing a course belonging to the same category improves.

There is going to be a lot more enhancement and research in this domain. With the world moving towards everything online, the scope of MOOCs only seems to be increasing which will result in more scope of improvement in this domain.

Refrences

[1] Chi Z, Zhang S, Shi L. Analysis and Prediction of MOOC Learners' Dropout Behavior. Applied Sciences. 2023 Jan 13;13(2):1068.

[2] Alshabandar R, Hussain A, Keight R, Khan W. Students performance prediction in online courses using machine learning algorithms. In2020 International Joint Conference on Neural Networks (IJCNN) 2020 Jul 19 (pp. 1-7). IEEE.

[3] Yimin Wen , Ye Tian, Boxi Wen, Qing Zhou, Guoyong Cai, and Shaozhong Liu (2020), Consideration of the Local Correlation of Learning Behaviors to Predict Dropouts from MOOCs.

[4] Wentao Li , Min Gao , Hua Li , Qingyu Xiong, Junhao Wen and Zhongfu Wu (2016), Dropout Prediction in MOOCs Using Behavior Features and Multi-view Semi-supervised Learning.

[5] Liang J, Li C, Zheng L. Machine learning application in MOOCs: Dropout prediction. In 2016 11th International conference on computer science & education (ICCSE) 2016 Aug 23 (pp. 52-57). IEEE.

[6] Lu X, Wang S, Huang J, Chen W, Yan Z. What decides the dropout in MOOCs?. InDatabase Systems for Advanced Applications: DASFAA 2017 International Workshops: BDMS, BDQM, SeCoP, and DMMOOC, Suzhou, China, March 27-30, 2017, Proceedings 22 2017 (pp. 316-327). Springer International Publishing.

[7] Feng W, Tang J, Liu TX. Understanding dropouts in MOOCs. In Proceedings of the AAAI Conference on Artificial Intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 517-524).

[8] Mourdi Y, Sadgal M, El Kabtane H, Berrada Fathi W. A machine learning-based methodology to predict learners' dropout, success or failure in MOOCs. International Journal of Web Information Systems. 2019 Oct 15;15(5):489-509.

[9] Deeva G, De Smedt J, De Koninck P, De Weerdt J. Dropout prediction in MOOCs: a comparison between process and sequence mining. In Business Process Management Workshops: BPM 2017 International Workshops, Barcelona, Spain, September 10-11, 2017, Revised Papers 15 2018 (pp. 243-255). Springer International Publishing.

[10] Jeon B, Park N, Bang S. Dropout prediction over weeks in MOOCs via interpretable multi-layer representation learning. Ar Xiv preprint arXiv:2002.01598. 2020 Feb 5.

[11] Earliest Predictor of Dropout in MOOCs: A Longitudinal Study of FutureLearn Courses, by Alexandra Ioana Cristea, Ahmed Alamri, Mizue Kayama, Craig Stewart, Mohammad Alshehri, Lei Shi, published in the year 2018.

[12] Monllao Olive D, Huynh DQ, Reynolds M, Dougiamas M, Wiese D. A supervised learning framework: Using assessment to identify students at risk of dropping out of a MOOC. Journal of Computing in Higher Education. Apr 2020 Apr; 32:9-26.

[13] Mourdi Youssef, Sadgal Mohammed, El Kabtane Hamada, Berrada Fathi Wafaa (2019), A predictive approach based on efficient feature selection and learning algorithms' competition: Case of learners' dropout in MOOCs.

[14] Hong B, Wei Z, Yang Y. Discovering learning behavior patterns to predict dropout in MOOC. In2017 12th International Conference on Computer Science and Education (ICCSE) 2017 Aug 22 (pp. 700-704). IEEE.

[15] Olive DM, Huynh DQ, Reynolds M, Dougiamas M, Wiese D. A quest for a one-size-fits-all neural network: early prediction of students at risk in online courses. IEEE Transactions on Learning Technologies. 2019 Apr 14;12(2):171-83.

[16] Fei M, Yeung DY. Temporal models for predicting student dropout in massive open online courses. In2015 IEEE international conference on data mining workshop (ICDMW) 2015 Nov 14 (pp. 256-263). IEEE.

[17] Qiu L, Liu Y, Hu Q, Liu Y. Student dropout prediction in massive open online courses by convolutional neural networks. Soft Computing. 2019 Oct; 23:10287-301.

[18] Gardner J, Yang Y, Baker RS, Brooks C. Modeling and Experimental Design for MOOC Dropout Prediction: A Replication Perspective. International Educational Data Mining Society. 2019 Jul.

[19] Whitehill J, Mohan K, Seaton D, Rosen Y, Tingley D. Delving deeper into MOOC student dropout prediction. arXiv preprint arXiv:1702.06404. 2017 Feb 21.

[20] Jha NI, Ghergulescu I, Moldovan AN. OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques. InCSEDU (2) 2019 May (pp. 154-164).

[21] Xing W, Du D. Dropout prediction in MOOCs: Using deep learning for personalized intervention. Journal of Educational Computing Research. 2019 Jun;57(3):547-70.

[22] Di Sun, Yueheng Mao, Junlei Du, Pengfei Xu, Qinhua Zheng, Hongtao Sun (2019), Deep Learning for Dropout Prediction in MOOCS.

[23] Alruwais N. Deep FM-Based Predictive Model for Student Dropout in Online Classes. IEEE Access. 2023 Sep 5.

[24] Liao J, Tang J, Zhao X. Course drop-out prediction on MOOC platform via clustering and tensor completion. Tsinghua Science and Technology. 2019 Mar 7;24(4):412-22.

[25] Link to the source of the dataset used for the experiment: http://moocdata.cn/data/user-activity#User%20Activity