Evolving Classification of Learners' Profiles using Machine Learning Techniques for Better Retention Rates in Massive Open Online Courses

MSc Research Project
MSc Data Analytics

Parth Patel
Student ID: x17164206

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Parth Patel
Student ID:	x17164206
Programme:	MSc Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	12/08/2019
Project Title:	Evolving Classification of Learners' Profiles using Machine Learning Techniques for Better Retention Rates in Massive Open Online Courses
Word Count:	5991
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	10th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Evolving Classification of Learners' Profiles using Machine Learning Techniques for Better Retention Rates in Massive Open Online Courses

Parth Patel
x17164206

Abstract

High dropout rates and low completion rates have been associated with Massive Open Online Courses (MOOCs) ever since its advent. The rigid structure, lack of feedback and interactivity provided by MOOCs imparts students with lack of motivation to complete the course. While many studies have been formulated to offer solutions to these problems, most of them require major design changes and are not cost effective. This paper aims to provide a design that can be used universally with any course and doesn't require design change in already existing MOOCs.

Objective: This study aims to build evolving classifier models, thatd classify students constantly based on their interactions with the courses materials.

Dataset: The dataset is obtained from Open University Learning Analytics, which contains details of students interaction with several courses materials and other demographic details of students.

Methodology: This research project compares 7 different classification algorithms, all implemented and evaluated in Python. The first occurrence of a sequence of 5-day data is trained on all the models, and the best performing model is then selected to continuously classify the students as the course progresses.

Results: This study find the Area Under ROC curve (AUC) score for every classifier. A score greater than 0.70 is considered a very strong course in studies in regard to predicting future behaviour changes, XGBoost outperformed all the algorithms, with AUC score of 0.73.

1 Introduction

Massive Open Online Courses, abbreviated as MOOCs, are talk of the decade. In 2018, there were 101 million students registered for more than eleven thousand courses[1]. But the MOOCs have been plagued with the problems of low retention of students. In the year 2018, just 3.13% of students completed their courses (Lederman, 2019). Many researches have been carried out to identify the cause of this problem, with researches not being able to single-out a cause. Many cite lack of motivation as the reason while no teacher-student interaction or less interactivity or response from the MOOCs are the reasons students fail to continue with their enrolled courses.

Various solutions have been proposed over the time to fix the problem that is low retention rates in MOOCs, but these studies have been performed in controlled group or

[1] https://www.classcentral.com/report/mooc-stats-2018/

have just been specific to single type of course. A design solution which can be universally applied to all the MOOCs would cater to this problem. This paper explores such works and proposes and implements one such solution; in order to increase students' motivation and reduce attrition rates.

1.1 Motivation:

This project aims to come up with a solution to improve the retention rates of the MOOCs, with a design solution that can be used for any course without requiring any major redesign in an already existing MOOC. An ideal solution would be cheap and not require human intervention; preserving the very purpose of MOOCs. While many researches have proposed methods thatd work towards improving the retention rates of the MOOCs, most of them are often associated with a major redesign or costlier alternatives; in an effort to make them more interactive, to increase the motivation of the students to continue with course. In one such effort Vaibhav and Gupta (2014), gamification of MOOCs was proposed and implemented. While this did prove that gamification helped with better completion rates, this required a major design change and was not cost effective. Even more, this solution is not universal: a new design would be required for every course.

1.2 Research Objective:

Motivation amongst students to complete the course and the level of interactivity provided by the MOOCs plays an important role towards completion of the courses students enroll for. If this is provided in a MOOC, chances are that a student feels obligated to complete a certain course. This is the aim of the research project; to come up with a design solution that provides frequent interaction to the student, in order provide them with impetus.

This research deals with data mining, advanced data processing to analyse and apply machine learning techniques on large sets of data and notifying students about their progress in the course frequently. Data mining is undertaken to extract the appropriate features from the huge dataset, classification algorithms are used to classify the students in the sequence of every 5-day interval. The aim of this research project is as follows:

- Preparing dataset in a manner which covers the temporal progress or interaction of the students. The time period of 5-day interval is considered for this project.
- Apply various classification algorithms on the first 5-days worth of data and evaluate the results obtain to determine the best performing algorithm.
- Train the remainder of the sets temporal data on the best performing model.
- If after any given 5-day interval theres a change in students behaviour, notify the student if the student is classified as at-risk.

The question of research here is: ***To what extent classification algorithms can be used to consistently evolve statuses of students in Virtual Learning Environments to increase their motivation and completion rates of Online Courses?***

Figure 1 outlines the general objective of this paper's solution. A learner will be consistently classified based on their interaction with the courses' materials.

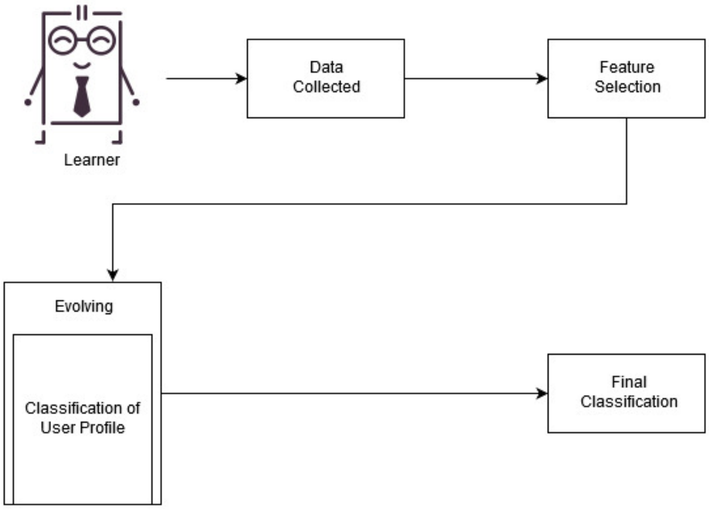


Figure 1: Proposed system for Better Feedback to Students

The rest of the paper structure is as follows: Section talks about the already existing related work in this sector, sector 3 talks about methodology being used for this research project. Section 4 and section 5 refers to the implementation and result. Finally, section 6 concludes this research

2 Related Work

2.1 Low Retention Rates in MOOCs: Causes

To fix the problem that is low completion rates of MOOCs and ever lower retention rates of the students, the first step is to identify what causes students to dropout from these online courses. Many researches have been undertaken to pinpoint the causation. This section reviews the work of such studies.

In the courses with lower audience, the problem of low completion rates still exists. In a research done by Hone and El Said (2016) on a small group of students from University of Cairo, the issue of low completion rates persisted. The research was conducted on a group of 379 students, with only 32% of them able to see the course through. When these participants were enquired via a questionnaire, many participants were of the opinion that better course interaction and extended support would help them to better stay on course with the course. They also established a correlation between constant feedback and improved rates of completion. In an another study done to determine the cause of this problem, Chaw and Tang (2019) used a scale; called Motivation and Engagement Scale in an effort to provide MOOCs designers will a solution that would make the courses more engaging. They collected data by using responses from the students and concluded that positive motivation had high correlation with engagement and eventually resulted in better completion rates.

In an another effort to relate the dismissal completion rates with different features of students' behaviour, Niu et al. (2018) considered 50 such features to discover the causation of low retention rates. They performed their research on more 1.5 million student click stream data to conclude that students are more likely to engage with the course if they are provided with more personal guidance and timely reminders about their interaction with the course. In their paper, they quote ten such features that have positive correlation with the completion rates of MOOCs and all those features are related to motivation and emotions of the students participating. In another study by Onah et al. (2014), they cited that the structure of courses and the quality of the materials used for presentation plays the crucial role in the engagement patterns and behaviour of the student. However, the major source of the problem is recognized as lack of support. In their experiment, they also devised that many students wish to learn at their own pace and even if they are not able to keep up with the course timeline, they would still want to continue the course. In conclusion, they had an opinion that if the MOOCs are too rigid and not adaptive in terms of providing interaction to the students, they don't feel motivated enough to continue.

To summarise, lack of motivation and support, not enough personal interaction and nonadaptiveness of the MOOCs drives the participants away from courses. Next section surveys solutions offered to fix these problems.

2.2 Improving Completion and Retention Rates: Solutions

Ever since the advent of the MOOCs, there have been complication related to low attrition rates. To overcome this, many solutions have been offered by various studies over the years. This section surveys those studies done to improve motivation and the interactions of the students undertaking online courses.

In an effort to increase the interactions of the students undertaking courses, various forms of redesign of the MOOCs have been proposed. In one such study done by Huang et al. (2019), game elements were introduced in MOOCs to study its effects in students' retention and interactions. They deduced that gamification helped participants create more network with other students. And if a student was associated with more networks, the same student had better engagement with courses' materials.

Another study done by Ortega-Arranz et al. (2019) on emphasised on reward based strategy to study its effect on students completion and retention rates. They deduced that students enrolled in courses which offered redeemable rewards were more likely to complete the courses. Also students were preferred first to complete the tasks which offered rewards rather than badges. Although more students participated in tasks which were gamified, the overall retention rate of the course did not increase significantly, as only courses with game elements and rewards were completed.

Use of narrative approach; such as use of storyline to impart knowledge to students and its impact on attrition rate was also studied in a study done by Pike and Gore (2018). They redesgined the course in a way that would include a narrative, whose progress was divided in a serialised sequence, in a form of a story. This made students to show up to next chapter of story, to find out what happened next. This design change correlated with high retention rates.

In a another effort done to understand the methods affecting the retention rates of MOOCs, Albelbisi et al. (2018) came out with their custom Template Approach with 3 dimensional features such as presage, process and the product. Presage had features pertaining to learners such as interactivity and more importantly motivation, while process had points such as assessment designs, behaviour patterns. They concluded that the learners' motivation plays a mighty role in better attrition rates.

To understand cause of dropouts and promote learners' persistence to complete course, Jung and Lee (2018) explored relationship between presence of a teacher, ease of use, level of engagement offered by the learning materials and completion rates of MOOCs. They conducted their research on 306 learners. They were very strongly able to conclude that if a MOOC offers a teacher, the student attrition rate was acceptable. They were of the opinion that the MOOCs should be designed in a way that would offer more learning support to the learners.

Imparting the sense of motivation in the students also leads to better retention rates as concluded in the study done by Xiong et al. (2015). The hypothesis proposed by them stated that motivation of the students predicted the retention rates and the motivation of the student is related to their engagement is the course. They had opinion for design changes in MOOCs in a way that would offer more interactivity to the students, thus increasing their engagement with the course and in-hand increasing the motivation of the students.

Most of the solutions offered by these studies works only if the course is redesigned or there's a presence of teacher. But this defeats the purpose of free-to-enroll MOOCs, as redesigning these courses would drive up the cost of these MOOCs and every other course would require a different approach to the solution.

Author	Outcome
Huang et al. (2019)	Gamification of MOOCs helps with better retention rates
Ortega-Arranz et al. (2019)	Reward based tasks had better completion rates
Pike and Gore (2018)	Narrative based approach design in MOOCs had better attrition rates
Albelbisi et al. (2018)	Students' motivation as well as product design matters
Jung and Lee (2018)	Presence of a teacher helps with better retention rates
Xiong et al. (2015)	Motivation is an important factor in students for them to continue with the course.

Table 1: Summarized Literature Review to Improve Retention Rates

2.3 Work Done on OULAD Dataset

The dataset being used for this research project has also featured on many other studies in the last 2 years; ever since its disclosure to public. This section surveys few of those recent researches.

In a research conducted by Haiyang et al. (2018), they used classification techniques to predict the dropout rates. They applied their techniques on the small part of dataset, and tried to predict whether the student would dropout or not early during the course. They were of the opinion that if students are informed earlier regarding their chances of failure, student would act accordingly to cut some slack. In an another study conducted by Rizvi et al. (2019), they tried to eastablish the relationship between the demographics of the students and their completion rates. They used Decision Trees to establish the relationships. They deduced that multiple deprivation, region and level of highest education achieved contributed most towards the completion rates of the MOOCs.

Hassan et al. (2019) made use of deep learning to predict the withdrawal of the students in virtual learning environment, again making use of the same dataset. Like in last surveyed research, they also made use of temporal sequential data to predict the probability of the dropout of the student, early in the course. They made use of interactional logs to leverage the deep learning technique such as Long Term Short Memory (LSTM) to achieve better performance than baseline machine learning models.

2.4 Conclusion

This section surveyed the relevant literature; first to pinpoint the causes of low retention in MOOCs, solution offered by other studies and work carried out on same dataset. While most cite lack of motivation and low interactivity provided by MOOCs as cause of low completion rates, the solutions offered by other researches often requires redesign of course or have more human presence to improve the interaction and motivation of the students. This research project provides with one more solution to provide more frequent feedback and interactions to students, in an effort to increase their motivation.

3 Methodology

This research focuses on the way to come up with a new business solution, a solution which will help provide the students of MOOCs with a more frequent, useful way of interaction. This is proposed to be done with the help of a notification system, which would inform whether they are on their paths to Pass the course or Fail, based on their interaction with the courses materials. The students will be classified every 5 days, and if their status changes, they will be sent a notification. This is proposed in order to make any MOOC more interactional, without requiring the human intervention or any major design change in existing MOOCs.

For this, Cross Industry Standard Process for Data Mining, also abbreviated as CRSIP-DM, will be applied.

3.1 Business Understanding:

As summarized in the literature survey above; the MOOCs are popular way of seeking education among the masses. But these MOOCs and the organisations providing them have been ridiculed with a serious problem: high withdrawal rates. These MOOCs have been facing lower retention rates, and even lower passing percentages. Many solutions have been proposed to improve upon these problems, with most requiring human interventions or major design overhaul; defeating the purpose of the businesses offering MOOCs, which is to keep them cheap and not have a requirement of a human presence.

This problem is solved using a two-way approach, both dependent on each other, applied over the temporal progress of any MOOC. These are:

1. To classify student every 5 days based on their interaction with the courses materials. The classification groups are Pass or Fail; to be done with the help of classification model already trained with the previous students interactional logs.

2. If a student is deemed to Fail based on any of the 5 days performance, send out a notification to the student informing about this.

While first step of classifying the students based on their interaction with the courses materials is the first most and important step, the second step of notifying the student of their progress and status is this research projects solution to make the MOOCs more interactive, providing the students with proper and frequent updates.

3.2 Data Acquisition:

The data used in this research comes from J. et al. (2017); which contains the general information about students partaking in the courses offered online, along with information generated as they progress in their courses. This information is their scores obtained in various assessments, and interaction logs in terms of total clicks done on any course content. The data is for seven different courses. Since this research focuses on the interactions done by the student on courses material and provide them updates on their status; this research will focus on total number of clicks done to classify students; along with their general information such as their previous education, number of attempts in any particular course, etc. The important features will be selected after performing couple of dimension reduction techniques.

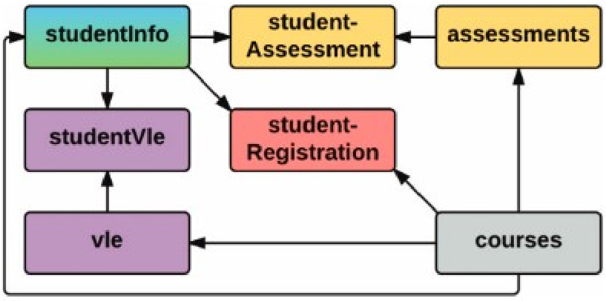


Figure 2: Structure of Data Files

Figure 2 demonstrates the structure of data, which further will be refined and processed to fit into a form much better suited for this research.

File	Description
studentInfo	Contains information pertaining to the students
vle	Contains information about the courses materials offered
Assessments	This contains the information about the assessments offered in the courses
Courses	This contains information about all the courses offered.
studentVle	This contains information on how the student interacts with VLE material.
studentAssessment	This contains information on how the student performed in assessments
studentRegsitration	This contains information about the registration details of the students.

Table 2: Brief Description of Data Files

Table 2 lists out the content of entity in the above entity relationship diagram, downloaded as comma-separated files from the source.

3.3 Data Preparation:

The research here focuses on three files to obtain the relationship between students interaction and their evolving behaviour, in terms of interaction with courses materials; as the course progress processes. These files are studentInfo, courses and studentVle. These files and their fields are described further below:

studentInfo: Contains the demographic information of the students participating in the courses. Also contains fields such as their student ID, gender, region, the highest level of education, index of multiple depravation, age group, the number of credits studied, whether the student is disabled or otherwise and finally the result obtained in the course.

courses: This file contains the available courses and their presentations along with the courses length in days. The presentation can either be B or J depending whether the course starts in February or October.

studentVle: This file contains the information about the students interactions with MOOCs Virtual Learning Environment (VLE) materials. The interactions are denoted as total number of clicks done on the VLE content. The filed date is field of major focus here, which lists the date in number of days the student interacted with the courses material, since the beginning of the presentation.

These three files are merged in a way to form a single dataset, containing students demographic details and their interactions with any given courses materials. The result of the student is also present in the file, which is the field used to determine the status of the student. The data will undergo additional preprocessing; before applying machine learning classification algorithms and notification system.

3.4 Feature Engineering:

Feature engineering correlates to converting the raw data into fields or features which can be used to increase the predictive power of the classification. In this research project, many such transformations will be done to the raw data, to make them more useful towards general predictiveness of the machine learning models. These engineering techniques will involve creating buckets, crossed value columns. These are further talked about during the data preprocessing stage of CRISP-DM.

3.5 Modelling

3.5.1 Naive Bayes

Naive Bayes is one of the most common and simple algorithms used for classification problems. It works on Bayes theorem, calculating conditional probability to determine certainty of occurrence of any event. According to an extensive research carried out by Mohanapriya and Lekha (2018a), Naive Bayes can deal with noisy and huge amount of data and it is also fast performing algorithm. However, the accuracy of a Naive Bayes model is not always up to the standards.

3.5.2 Decision Tree

Decision tree is also a good choice for classification problems, and according to a research published by Song and Lu (2015), the advantages of using a decision tree for classification problems such as; simplification of relationships which are complex, no assumptions about data distribution and no affect of outliers make this algorithm a perfect choice for our research.

3.5.3 k-Nearest Neighbour

K-nearest neighnour or k-NN is a non-parametric technique used for regression as well as classification problems. Mohanapriya and Lekha (2018b) carried out researches to compare k-NN with decision tree, they concluded that k-NN is effective for small data, but tends to underperform under the influence of noise.

3.5.4 Bagging with Decision Tree

To improve the performance of decision tree and to reduce the variance, bagging is performed. Bagging takes subsets of data, always with replacement and train the model with each individual data sets. This helps the model to learn more and eventually increase the performance of model.

3.5.5 Logistic Regression

Logistic regression is the simplest algorithm that this research project will make use of. Since the problem here is that of a binary classification; training a logistic regression algorithm model makes sense. As it is fast and doesn't have nearly as much computational requirements like the other algorithms, logistic regression will serve as the baseline model in this research project.

3.5.6 Neural Network

Neural networks are being used actively in classification problems as they can learn quickly; if right choices are made in regards to hyper-parameter tuning and activation functions. For this research project, neural network with three hidden layers will be modelled.

3.5.7 XGBoost Classifier

XGBoost can be used to regression or classification problems. Since XGBoost is a combination of ensemble trees and gradient descent learning. Since a new tree is added each time based on maximum depth specified, a new XGBoost model almost always perform better than old model. Since this research requires such a robust algorithm, XGBoost is fine choice here.

3.6 Evaluation:

Since the classification of the students status will be done with the help of various classification algorithms introduced above, the appropriate way to figure out how well our models perform will be how effectively these models will be able tell the current status of the students, based on their interaction with the courses materials. There are various metrics to evaluate classification algorithms; such as classification accuracy, logarithmic loss, area under ROC Curve.

4 Implementation

The implementation of the above proposed solution follows a pipeline written in python, which starts from reading the data files using Pandas, followed by merging and filtering for a specific course, feature engineering, data preprocessing, data preparation and applying machine learning algorithms. Figure 3 shows the flow of implementation:

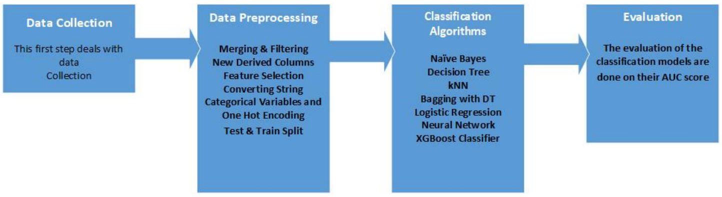


Figure 3: Implementation Work Flow

4.1 Data Collection:

The data files were downloaded from J. et al. (2017), as comma separated files, abbreviated as CSV. The files used for this project are students information, students interactions with courses material and details about the courses. The data is read into Pandas dataframe using python for further data preprocessing.

4.2 Data Preprocessing:

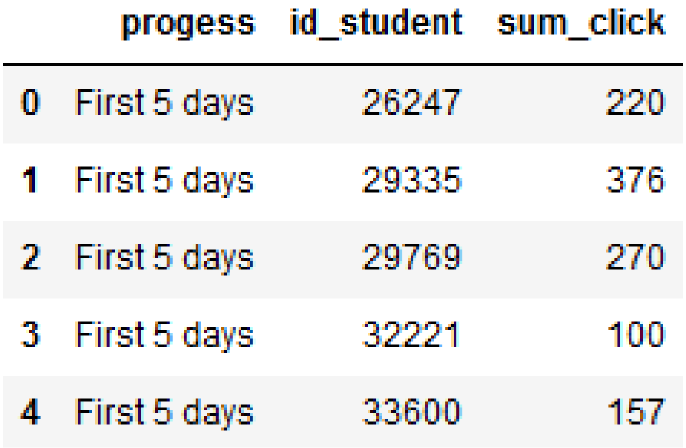
Data preparation; for transforming data from its raw form into a form which is ideal for machine learning algorithm models, is the important step for any machine learning project. The various transformation steps that the raw data undergoes are explained in this section:

4.2.1 Merging and Filtering:

The 3 data files are merged on similar fields amongst themselves. The file studentInfo and studentVle are merged on the key student_id. The resulting file is then merged with the file courses, on the key course_id. The resulting file contains details about every student undertaking several courses being offered, their demographic details, the interactional logs in terms of total number of clicks and their result. The resulting file is then filtered on course and presentation ID, in a way that a single file just contains details about a single course and its presentation period.

4.2.2 Derived Columns:

Since this research project focuses on classifying the evolving behaviour of the students every 5 days into course, it is imperative to derive a column thatd tag the progress of students, every 5 days. Also, one of the aims is to figure out the students interaction with courses materials every 5 days, so the total number of clicks done by each student every 5 days would be summed, as a way to indicate students interaction every 5 days. Figure 4 shows how derived columns would look like.



	progess	id_student	sum_click
0	First 5 days	26247	220
1	First 5 days	29335	376
2	First 5 days	29769	270
3	First 5 days	32221	100
4	First 5 days	33600	157

Figure 4: Progress: A New Derived Column

4.2.3 Feature Selection:

Feature selection refers to selecting the best features from the dataset that directly affects towards predicting the target variable, in our case that is the result of the students. The variables which have the strongest relationship with the output variable will be used to train the machine learning models. The feature selection was done with the help of feature importance, implemented with the help of ExtraTreeClassifier included in a Python package sklearn Pedregosa et al. (2011).

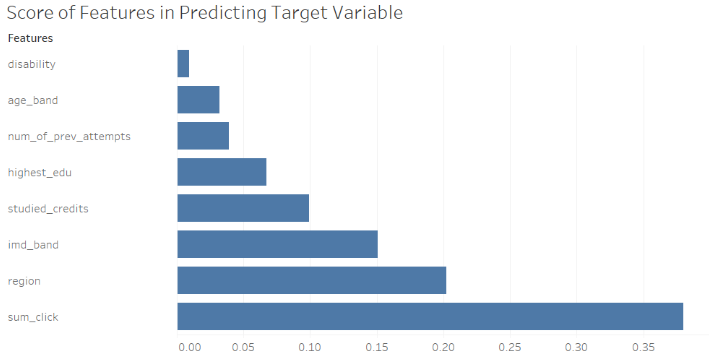


Figure 5: Features and their scores.

As seen clearly from the above graph in Figure 5, the field sum_click has the strongest relationship with the output variable; final_result. The other fields such as region the student belongs to, their depravation index, number of studied credits and their highest level of education also affects the outcome of the target variable.

4.2.4 Converting String Categorical Variables and One Hot Encoding:

Once the appropriate features were figured out, the remaining step was to prepare data for in a way that a machine learning model can understand. Since most of the fields are string and are nominal, they were converted to integers with the help of a method cat.codes, which essentially converts all the categorical variables into codes, ranging from 0 to n_categories–1, sorted alphabetically. But this creates a new problem for the machine learning model, as these models starts to associate the integers with their numerical power. This problem was solved with the help of One Hot Encoding of these variables, which essentially creates a new column for each category in every categorical variable, converting the dataframe in a manner that is understood by the machine learning models.

4.2.5 Test and Train Split:

The final prepped dataset was then split into test and train, with training data occupying 80% of the data. The model was tested on 20% of the remaining data.

4.3 Classification Algorithms:

Various types of classification algorithms are used for classification of the students based on their interaction with the courses materials. Since this project proposes classification of students every 5 other days based on their behaviour, the best performing model for the first 5 days will be selected for continued classification of the students in the upcoming days. The models were chosen based on their performance, evaluated with the help of various scores such as model accuracy, precision and recall scores and also area under ROC (receiver operating characteristic curve), abbreviated as AUC scores. AUC scores here are deemed as essential to determine a models performance. The models used here are Nave Bayes, Decision Tree, K-Nearest Neigbhour (kNN), Bagging with Decision Tree, Logistic Regression, Neural Network, Random Forest, Support Vector Machine (SVM) and XGBoost Classifier.

5 Evaluation

The evaluation of these algorithms was done with the help of scores as discussed above. The following Table 3 talks about all the algorithms ran and their AUC scores in the first 5 days of the course FFF:

Metrics	Models						
	Nave Bayes	Decision Tree	kNN	Baggingwith Decision Tree	Logistic Regression	Neural Network	XGBoost Classifier
Accuracy	0.59	0.58	0.72	0.66	0.67	0.68	0.72
Precision	0.68	0.61	0.74	0.64	0.71	0.61	0.74
Recall	0.49	0.65	0.75	0.72	0.64	0.65	0.73
AUC	0.60	0.62	0.72	0.66	0.67	0.64	0.73

Table 3: Models and their metrics scores

The above algorithms, except for Bagging with Decision Trees, were trained with stratified K folds, were the number of K was kept at 5. Stratified K fold was chosen as it allows preservation of the samples for each class, eliminating the need for balancing the dataset. The best performing model based on data for first 5 days was XGBoost Classifier. According to a study done by Rice and Harris (2005), for researches based on prediction of future behaviour prediction, an AUC score above 0.70 is considered a strong score. This fits exactly with the requirements of this research project. XGBoost will be used to train models with the further data as the course duration progresses.

The metrics on which these algorithms are evaluated are accuracy, precision, recall and Area Under ROC (Receiver Operating Characteristic) scores. Since, this being a classification problem, the point of interest here is how accurately the models were able to classify the students each time. These metrics are further talked about and how they can be used as a measure to evaluate the performance of the models.

Model accuracy: Model accuracy is one of the simple methods to evaluate classification models. The formula for accuracy is defined as:

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Also for binary classification, such is our case, the following formula also works to determine the model accuracy:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where TP stands for True Positive, TN for True Negative, FP for False Positive and FN for False Negative.

Precision: Precision score is what proportion of positive cases were identified correctly. The formula for precision is given as:

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Where TP stands for True Positive and FP for False Positive.

Recall: Recall is what proportion of actual positives were correctly identified.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Where TP stands for True Positive and FN for False Negative.

AUC Score: Area Under ROC curve measures the 2D area under curve of ROC. ROC Curve is obtained by plotting two parameters which are: True Positive Rate and True Negative Rate. True Positive Rate is same as recall, while False Positive Rate or FPR is given as:

$$FPR = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Where FP stands for False Positive and FN for True Negative.

A model with AUC score of 1 is considered to have no wrong predictions while a model with AUC score of 0 is considered to have all the predictions wrong.

Since AUC score considers all the thresholds of classification, it is an obvious choice here to judge the models based on their AUC Score. XGBoost had the best AUC score amongst all the models.

5.1 Experiment 1. Nave Bayes

Nave Bayes was ran using 5 stratified folds of the preprocessed dataset. The performance of this model was not too good for when trained and tested on the data for first 5 days. But this was expected as Naive Bayes as its assumption independence often correlates to models performance.

5.2 Experiment 2. Decision Tree

Decision tree performed better that Nave Bayes, but it still failed to hit the AUC score of above 0.70, a score considered strong in studies regarding future behaviour predictions. To further increase the performance of decision tree, the data was split in 10 folds, but the performance of the model did not change significantly.

5.3 Experiment 3. kNN

kNN had the second-best performance of all the models trained on the first 5 days data. The AUC score of 0.72 is strong considering the nature of the research. For this, the data was again prepared in stratified 5 folds and then mean of all the scores were considered.

5.4 Experiment 4. Bagging with Decision Tree

Since the performance of decision tree was dismissible, bagging method was used to boost the performance of the decision tree algorithm. The scores improved significantly, although the AUC score was still below 0.70.

5.5 Experiment 5. Logistic Regression

Logistic regression was selected as this being a binary classification problem. The logistic regression model performed quite good for being not so complex algorithm, although the AUC score was still below 0.70. These scores were often on a model trained with 5 fold stratified datasets.

5.6 Experiment 6. Neural Network

Neural Network, with three layers of activation functions: rectified linear unit, rectified linear unit and sigmoid, performed on average with other models, failing to match the performance provided by models such as XGBoost Classifier, but still had decent AUC score of 0.64.

5.7 Experiment 7. XGBoost Classifier

XGBoost outperformed all the models, with the AUC score of 0.73. The 5 stratified folds were again used to train the model, and the mean scores of all the evaluation metric were above 0.70, making XGBoost an ideal choice to train models for data of every other 5 days. For simplicity, the data for up to 25 days were trained with XGBoost models, with interval of 5 days.

The AUC scores when the data for different intervals was trained using XGBoost classifier were consistent; always close to 0.70. This made XGBoost as the algorithm of choice while building the library of behaviour models and implementing the test cases.

6 Test Cases:

The test cases are designed in a way to test the XGBoost models created for specific intervals down the line as the course progresses.

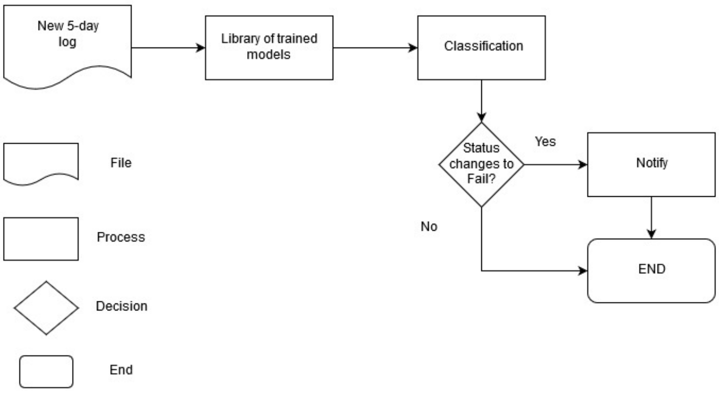


Figure 6: Design Flow for Test Cases

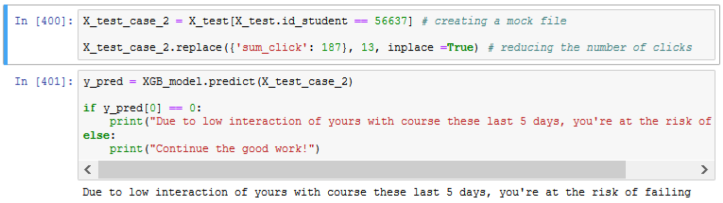
In Figure 6, the general working of test cases is shown. Whenever a log file for a specific 5-day interval is created for a student, it looks for a model already trained on previous students' data for that interval. The model tries to reclassify the student based on their current interaction with the courses' material. If the status of student changes to "Fail", the student is notified.

6.1 Test Case 1:

For the first test case, a student is selected whose final result is Fail. For the model trained for days 11-15, a mock log file is created, with changes to total number of clicks done. The value is changed to be on par with students whose final result is Pass. When such mock logfile was passed through the already trained model, the model was accurately able to change the status of student to Pass, thus considering the changed behaviour of the student with regards to the interaction and evolving the students profile successfully.

6.2 Test Case 2:

Here, the opposite of the first test case is done. The interaction of the student with result Pass was made low. The value of the field sum click was changed to match with those whose result is Fail. The model was able to successfully change the status of the student to Fail and was also able to send out a notification as well.



```
In [400]: X_test_case_2 = X_test[X_test.id_student == 56637] # creating a mock file

          X_test_case_2.replace({'sum_click': 187}, 13, inplace =True) # reducing the number of clicks

In [401]: y_pred = XGB_model.predict(X_test_case_2)

          if y_pred[0] == 0:
              print("Due to low interaction of yours with course these last 5 days, you're at the risk of
          else:
              print("Continue the good work!")

          Due to low interaction of yours with course these last 5 days, you're at the risk of failing
```

Figure 7: A previously Passing Student reclassified as "Fail" due to low activity

6.3 Discussion:

In this project, an ideal classification algorithm was sought based on its performance on first 5 days data. The algorithm with best metrics and performance was selected for further classification of the students as the course progresses.

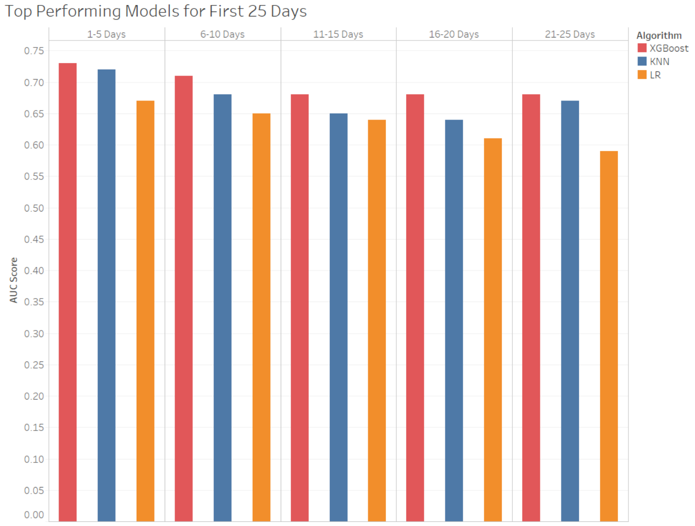


Figure 8: A previously Passing Student reclassified as ”Fail” due to low activity

The graph in Figure 8 depicts the performance of 3 models for first 25 days, in an interval of 5 days. The models are XGBoost, k-NN and Logistic Regression.

XGBoost Classifier had the best AUC score amongst all the other classifiers. AUC score of 0.70 was required because of the nature of this research i.e.; predicting the future of the students based on their interactions with the courses materials. The two test cases formed above also contributed towards the credibility of XGBoost classifier as an ideal classification algorithm for classifying students as the course progresses and their interaction with courses material changes. Naive Bayes performed not too good, but that was expected from a model which makes too much assumption about predicting class. Bagging with decision tree obviously had better performance than Decision Tree, as bagging was done with 100 such trees to obtain the optimal performance from the model. There was a difference of 0.08% in the accuracy score between decision tree and bagging with decision tree.

k-NN was the second best performing model after XGBoost. The model was trained with 5 stratified folds, thus helping mantain the balance between both the classes in each of the 5 sets of training data. The AUC score obtained was 0.72, just 0.01 less than XGBoost.

XGBoost was the best performing model, and to find the best hyper-parameter for it, grid search of parameters was performed to figure out the best parameter. These parameters were max_depth, learning rate and number of folds. The XGBoost was retrained by retuning these parameters to obtain even better performance.

The OULAD dataset was used before by many other researches, many working towards predicting the dropout or failure rates of the students undertaking these courses. In a research published by Hussain et al. (2018), lack of motivation is cited as the major reason for lack of the students continuation with the courses. While they had better classification scores, they performed their research on the entirety of data, not considering the evolving changes in the students behaviour, thus neglecting the temporal changes. Another research by AL-Shabandar et al. (2018) , although considering more factors to predict students dropout rates, neglects the temporal behaviour changes in students.

This is the main objective of this research paper, to not just classify the at-risk students based on their performance throughout the courses duration, but to also considering evolving changes in their interactions with course and notify them appropriate whenever theyre at-risk of failing. This is proposed to make the course more interactive, without requiring any major overhaul to the existing design or requiring any human intervention as well.

7 Conclusion and Future Work

This research project aimed at coming up with a way to continuously classify students based on their changing interactions with the courses' materials. To achieve this, the classification model was trained in a sequence of 5-day intervals. Several classification models were trained on the first 5 days' data, but XGBoost was used for further classification of students as course progresses; as it outperformed in comparison to other models. If the students' status changes from "Pass" to "Fail" due to low activity on MOOCs, theyd be classified accordingly and notified. This is aimed towards making the MOOCs as whole more interactive and personal.

This was done to in an effort to increase the motivation of students to complete the courses they enroll for, if they are consistently notified if they dont interact much, or even if their behaviour has changed which puts them at the risk of failing. While there's no way to know how much this method will help impart the sense of motivation in these students, the classification model performed well to change the status of at-risk students and notify them on timely basis to minimise the risk of dropout.

One of the future work of this would be built upon this, and to assess how much this solution helps with increasing the motivation of the students to complete course they have registered for. Also, this research just considers the interactional logs of the students to determine the status of the students. In future, more factors such as assessment scores, assignment submission would be consider to better classify the students.

References

AL-Shabandar, 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, pp. 1–8.

Albelbisi, N., Yusop, F. D. and Salleh, U. K. M. (2018). Mapping the factors influencing success of massive open online courses (mooc) in higher education, *Eurasia Journal of Mathematics, Science and Technology Education* **14**(7): 2995–3012.
URL: *http://dx.doi.org/10.29333/ejmste/91486*

Chaw, L. Y. and Tang, C. M. (2019). Driving high inclination to complete massive open online courses (moocs): Motivation and engagement factors for learners., *Electronic Journal of e-Learning* **17**(2): 118–130.

Haiyang, L., Wang, Z., Benachour, P. and Tubman, P. (2018). A time series classification method for behaviour-based dropout prediction, *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, pp. 191–195.

Hassan, S.-U., Waheed, H., Aljohani, N. R., Ali, M., Ventura, S. and Herrera, F. (2019). Virtual learning environment to predict withdrawal by leveraging deep learning, *International Journal of Intelligent Systems* **34**(8): 1935–1952.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/int.22129*

Hone, K. S. and El Said, G. R. (2016). Exploring the factors affecting MOOC retention: A survey study, *Computer Education* **98**: 157–168.

Huang, B., Hwang, G.-J., Hew, K. F. and Warning, P. (2019). Effects of gamification on students online interactive patterns and peer-feedback, *Distance Education* **0**(0): 1–30.

Hussain, M., Zhu, W., Zhang, W. and Abidi, R. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores, *Computational Intelligence and Neuroscience* **2018**: 1–21.

J., K., M., H. and Z., Z. (2017). Open university learning analytics dataset.

Jung, Y. and Lee, J. (2018). Learning engagement and persistence in massive open online courses (moocs), *Computers Education* **122**.

Mohanapriya, M. and Lekha, J. (2018a). Comparative study between decision tree and knn of data mining classification technique, *Journal of Physics: Conference Series* **1142**: 012011.

Mohanapriya, M. and Lekha, J. (2018b). Comparative study between decision tree and knn of data mining classification technique, *Journal of Physics: Conference Series* **1142**: 012011.

Niu, Z., Li, W., Yan, X. and Wu, N. (2018). Exploring causes for the dropout on massive open online courses, pp. 47–52.
URL: *http://doi.acm.org/10.1145/3210713.3210727*

Onah, D. F. O., Sinclair, J. and Boyatt, R. (2014). Dropout rates of massive open online courses : behavioural patterns, *EDULEARN14 Proceedings* **46**(1): 5825–5834.

Ortega-Arranz, A., Bote-Lorenzo, M., Asensio-Prez, J., Martnez-Mons, A., Gmez-Snchez, E. and Dimitriadis, Y. (2019). To reward and beyond: Analyzing the effect of reward-based strategies in a mooc, *Computers Education* p. 103639.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**: 2825–2830.

Pike, G. and Gore, H. (2018). The challenges of massive open online courses (moocs), *Creativity and Critique in Online Learning: Exploring and Examining Innovations in Online Pedagogy* pp. 149–168.

Rice, M. and Harris, G. (2005). Rice me, harris gtcomparing effect sizes in follow-up studies: Roc area, cohen's d, and r. law hum behav 29: 615-620, *Law and human behavior* **29**: 615–20.

Rizvi, S., Rienties, B. and Khoja, S. A. (2019). The role of demographics in online learning; a decision tree based approach, *Computers Education* **137**: 32 – 47.
URL: *http://www.sciencedirect.com/science/article/pii/S0360131519300818*

Song, Y.-Y. and Lu, Y. (2015). Decision tree methods: applications for classification and prediction, *Shanghai archives of psychiatry*.

Vaibhav, A. and Gupta, P. (2014). Gamification of moocs for increasing user engagement.

Xiong, Y., Li, H., Kornhaber, M. L., Suen, H. K., Pursel, B. and Goins, D. D. (2015). Examining the relations among student motivation, engagement, and retention in a mooc: A structural equation modeling approach, *Global Education Review* **2**: 23–33.