information

MDPI

Article

Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model

Sheran Dass [1], Kevin Gary [1,*] and James Cunningham [2]

[1] School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85251, USA; sdomini1@asu.edu

[2] EdPlus at Arizona State University, Tempe, AZ 85251, USA; Jim.Cunningham@asu.edu

* Correspondence: kgary@email.asu.edu

Abstract: A significant problem in Massive Open Online Courses (MOOCs) is the high rate of student dropout in these courses. An effective student dropout prediction model of MOOC courses can identify the factors responsible and provide insight on how to initiate interventions to increase student success in a MOOC. Different features and various approaches are available for the prediction of student dropout in MOOC courses. In this paper, the data derived from a self-paced math course, College Algebra and Problem Solving, offered on the MOOC platform Open edX partnering with Arizona State University (ASU) from 2016 to 2020 is considered. This paper presents a model to predict the dropout of students from a MOOC course given a set of features engineered from student daily learning progress. The Random Forest Model technique in Machine Learning (ML) is used in the prediction and is evaluated using validation metrics including accuracy, precision, recall, F1-score, Area Under the Curve (AUC), and Receiver Operating Characteristic (ROC) curve. The model developed can predict the dropout or continuation of students on any given day in the MOOC course with an accuracy of 87.5%, AUC of 94.5%, precision of 88%, recall of 87.5%, and F1-score of 87.5%, respectively. The contributing features and interactions were explained using Shapely values for the prediction of the model.

Keywords: prediction; dropout; MOOC; random forest; AUC; ROC; SHAP

check for updates

Citation: Dass, S.; Gary, K.; Cunningham, J. Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model. *Information* **2021**, *12*, 476. https://doi.org/10.3390/info12110476

Academic Editor: Gennady Agre

Received: 16 August 2021
Accepted: 4 November 2021
Published: 17 November 2021

1. Introduction

Massive Open Online Courses (MOOCs) are (typically) free, Web-based courses available to learners globally and have the capacity to transform education by fostering the accessibility and reach of education to large numbers of people [1]. They have gained importance owing to their flexibility [2] and world-class educational resources [3]. ASUx®, Coursera®, and Khan Academy® are some examples of popular MOOC providers. Since 2012, MOOC offerings have increased at top Universities [4]. Investigations undertaken by such institutions indicate that the use of MOOCs attracts many participants towards engagement in the space of courses offered due to the removal of financial, geographical, and educational barriers [4].

However, despite the potential benefits of MOOCs, the rate of students who drop out of courses has been typically very high [5–7]. Recent reports also show that the completion rate in MOOCs is very low compared to the number of those enrolled in these courses [8]; hence, the prediction of the student's dropout in MOOCs is essential [9]. Even though there are many reports on prediction, there is no prediction based on the features in machine learning (ML) using random forest (RF). The contribution of this paper is a prediction model of students' dropout in a MOOC for an entry-level science, technology, engineering, and mathematics (STEM) course using RF. While this model may be improved, we believe it is a valuable step to understand feature interaction and has applicability to similarly framed STEM MOOCs.

This paper is focused on predicting the dropout of students from MOOC with the help of ML by the application of RF using features that have not been used before. Two research questions are raised concerning this context:

> *RQ 1: What are the features of changes in learning progression that are associated with students who drop out of a MOOC course?*

> *RQ 2: Given a set of features of changes in the learning progression of a student on a day of consideration, can we predict the day of dropout of a student in a MOOC course?*

These research questions are of great significance because of the following reasons:

- ➢ Predicting when (the day) a student may drop out of the MOOC course helps in designing a targeted intervention that can bring the student back into the course.
- ➢ Many self-paced courses use Knowledge Space Theory, and this research could be extended to such courses.
- ➢ MOOC courses offering college credit, such as the one considered for this research, where students drop out would be interested in addressing this problem.

2. Related Work

Educational Data Mining (EDM) is the application of data mining techniques to educational data to obtain solutions to problems in the field of education [10]. EDM engrosses the use of statistics, visualization, and machine learning techniques for the assessment and evaluation of educational data [11]. Some of the EDM applications include the formulation of e-learning systems [10,12], clustering educational data [13,14], and making predictions of student performance [11,14–16].

Learning Analytics (LA) is an emerging field of research that intends to improve the quality of education [17,18]. There are various techniques exploited by researchers in LA, like Web analytics, artificial intelligence, and social network analysis [17]. The key feature of LA is its capacity to evaluate actionable data in a more objective way [18,19]. Although many works have been reported in the literature to analyze the learner performance in the e-learning environment, it is still challenging to construct predictive models for MOOCs [4].

Dropout in MOOCs refers to the event of students failing to complete the course [20]. Even though there are a great deal of reports on the prediction of student dropout in MOOC, it remains an important problem in this research area [9]. One of the reasons for this problem remaining important despite a decade-plus of MOOC offerings is that there has been no universal technique to predict student dropout that can be applied to multiple courses.

2.1. Feature Engineering

Feature engineering is emerging as an important technique. The incorporation of features, including test grades, within a course could prove to be a useful and effective solution to the prediction problem in EDM [5].

Several studies aim at evaluating features from learners' online activities [21–23], but few papers also use demographic features [24,25]. Typically, the features considered for analysis include study time, study duration, content type, and features derived from social interactions, but the emergence of the online learning platform as a stable and interactive platform transformed the features to assessment scores, assignment scores, clickstream analysis, online forum interaction, and location for the analysis process [26]. The selection and identification of significant features are some of the challenges for researchers due to diversity in platforms including MOOCs.

The role of demographic features has been analyzed on student rate of retention [27–30]. For example, [31] examined approximately 120 variables, including educational background, clickstream data, assessment scores, entry test scores, and learning personality data, to analyze impact on student performance. Even though most of the studies focus on finding the impact of key features on students' performance, there are studies that concentrate on early prediction, intervention, learned support, and appropriate feedback to guide and prevent student dropout [32–36].

Tal et al. [37] showed that students performing well demonstrate better engagement compared to the students of poor performance. Clickstream features, including the online engagement of students, are more accurate, objective, and comprehensive than self-reported data in measuring student's learning behavior [38]. Clickstream data are discreet and do not require a student's full attention, as they can be collected effortlessly without interfering with students' learning processes [39]. Although most of the studies try to investigate the relationship between clickstream data and students' online engagements (e.g., [40,41]), very few studies have gone one step further to enable instructors to know how and when to intercede with students at the optimal time [42–45].

2.2. Machine Learning

Creating predictive and flexible models that can adapt and/or adjust in different learning environments is a significant challenge. Limitations include the presence of various course structures, different instructional designs, and diverse online platforms [46]. Machine Learning (ML) is a proficient technique that can be applied to Learning Analytics with the ability to discover hidden patterns of student interaction in MOOCs with advantages over conventional forms of statistical analysis [4,17,19,47,48]. ML algorithms, such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees (DT), RF, Deep Learning (DL), and Recurrent Neural Network (RNN), have been applied to the prediction of students learning performance at the end of the course [49].

A recent prediction of student dropout in online courses using a time-series clustering approach was developed with better accuracy than the conventional aggregation method [50]. Gokhan et al. [51] developed another early warning system that employed students' eBook reading data to predict students' performance for academic failure. Among them, the best predictive model was selected based on the accuracy/Kappa metric [52,53], and the RF, J48, and Naïve Bayes (NB) showed better performance compared to others.

Students who have a high possibility of failure were analyzed by four ML algorithms for the early identification of their performance. Among them, the Support Vector Machine (SVM) was the most effective algorithm in the earlier identification of students. The accuracy was found to be 83%. Moreover, the preprocessing of data was found to be important in increasing the performance of ML algorithms [54]. Earlier studies describe the development of predictive models, but many challenges limit their application to a specific learning platform. Creating predictive and flexible models that can adapt and/or adjust in the different learning environment is a significant challenge. The limitations were the presence of various course structures, different instructional designs, and diverse online platforms [46].

Recently, researchers have used both statistical and predictive models to explore in a large repository, including formal and informal educational settings [47,48]. Alberto et al. [55] reported that the multi-view genetic programming approach to develop classification rules for the analysis of students learning behavior to predict their academic performance and trigger alerts at the optimal time to encourage at-risk students to improve their study performance. Logistic regression was also employed to identify students' dropout in an e-learning course [56]. This technique showed a higher performance score in validation including precision, recall, specificity, and accuracy than feed-forward neural network (FFNN), Support Vector Machine (SVM), a system for educational data mining (SEDM), and Probabilistic Ensemble Simplified Fuzzy Adaptive Resonance Theory Mapping (PESFAM) techniques.

Knowledge discovery in databases (KDD) was employed to mine information that may enable teachers in finding the interaction of students with e-learning systems [12]. A Decision Tree (DT) algorithm was used [57] to establish significant features that assist MOOC learners and designers in developing course content, course design, and delivery. Various data mining techniques were applied to three MOOC datasets to evaluate the in-course behavior of the online students. The authors claim that the models used could be beneficial in the prediction of significant features to lower the attrition rate.

These studies help in the prediction of student performance, including dropout rate; however, none of these studies predict students at-risk of dropout at a different stage of a course. Further, there is no study on the prediction of the dropout of students using RF with the features identified in this research. Hence, we report the RF model with features including average, standard deviation, variance, skew, kurtosis, moving average, overall trajectory, and final trajectory.

3. Data Description and Methodology

3.1. Data Description

The data from the self-paced math course College Algebra and Problem Solving offered on the MOOC platform Open edX offered by EdPlus at Arizona State University (ASU) from 2016 to 2020 was considered. Restrictions apply to the availability of these data. Data were obtained from EdPlus and are available from the authors with the permission of EdPlus. Additionally, this data cannot be made publicly available because it is private student data protected under the Family Educational and Privacy Act (FERPA). The work in this study is covered under ASU Knowledge Enterprise Development IRB titled *Learner Effects in ALEKS*, STUDY00007974.

The student demographic data were analyzed to get an idea of the background of the students, and such a description helps us in understanding the impact of this research. The distribution of the students in this course is shown in Table 1.

Table 1. Distribution of Students in the Course.

Class	Number of Students	Percentage
Complete	396	12.50%
Dropout	2776	87.50%

From Table 1, we see that out of the 3172 students in the course, only 396 students completed the course, while 2776 students dropped out of the course. This problem of dropout is seen in this course, and research has shown that it is very prevalent in MOOCs. We also looked at the demographic distribution by age, gender, and ethnicity (see Appendix A), and while we found primarily white, more male than female, and mostly 20 year-old learners, we did not detect any bias based on these moderators.

Our first prediction model attempted to apply a clustering approach utilizing the process suggested by [58]. Feature identification followed the work of [59], who performed a k-means clustering on a small EDM dataset to identify detrimental behavior to learning, which helps discover the relationship between behavioral interventions and learners' requirements. K-means clustering was then performed using Lloyd's algorithm [60]. This initial attempt yielded poor accuracy (22%) and led us to examine our model and what was occurring. The key insight was realizing the temporal aspect of rate-of-change in learning as the key predictor instead of simply clustering based on whether the student did indeed drop out of the course. We re-evaluated the data and applied a Random Forest (RF) classifier, as described next.

3.2. Methodology

The methodology is organized into three parts data handling, machine learning modeling, and model evaluation. The flow of the methods is explained in Figure 1.

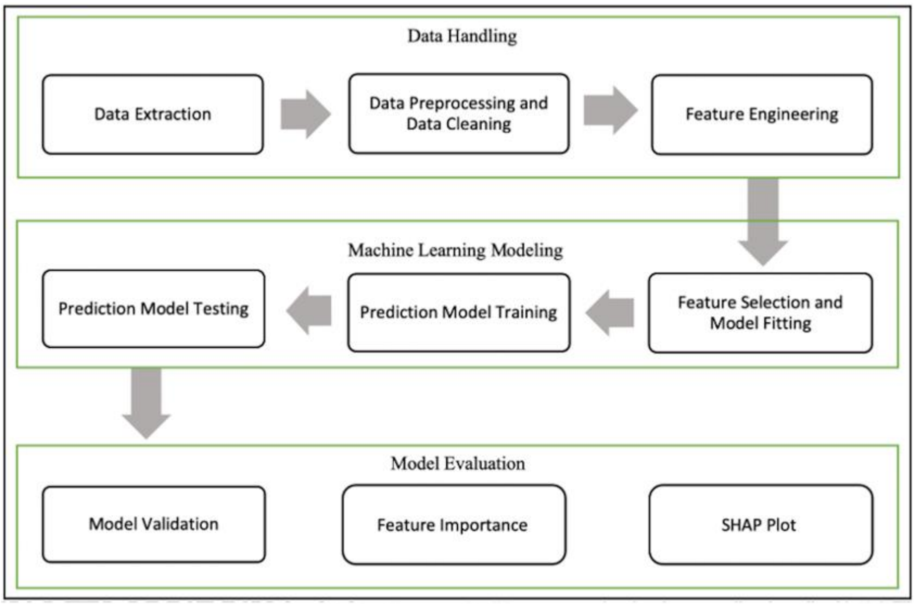

Figure 1. Methodology. The three components are data handling, machine learning modeling, and model evaluation. The steps within each component are explained in this section.

3.2.1. Data Handling

Data handling is done by the standard process of KDD [61]. The data handling involves three steps: data extraction, data preprocessing, and data cleaning and feature engineering.

3.2.2. Data Extraction

Data extraction involves extracting data from the data source by using data mining methods. The ALEKS platform has an API, which is used to access the data in the SQL Database. The data selection method was used to select query on the data source [62] from the SQL database. Once the data selection is performed, the queried table is stored as a comma-separated values (CSV) file. Jupyter notebook was used to access CSV file and processed by the python programming language.

The total number of participants in this course is 3172, and each has some level of activity after their Initial Knowledge Check (IKC). The IKC is a proficiency test conducted at the beginning of the course for all students to assess their current knowledge. Moreover, the ALEKS system adaptively designs the students' knowledge domain based on the IKC and progresses from their existing knowledge space.

3.2.3. Data Preprocessing and Data Cleaning

Data preprocessing is the process of extracting data needed for the machine learning model from the target data. The features of change in student learning that are associated with student dropout were found and used to predict this event. The extracted dataset has data as a time series with no primary keys. The grouping was performed based on student identity, which gives structure to the raw dataset [62]. Once grouped, the learning progression data for each student from the target data can be extracted and stored as a table called the preprocessed data.

A student in ALEKS [63,64] goes through a knowledge check after every topic or every 120 min in the platform, based on whichever event happens first. If they clear the knowledge check, then they are recorded as mastering the topics assessed, and if they do not clear, then they are recorded as not mastering topics (even if they mastered them before). These topics mastered serve as the measure of progress to a student, and when they cover 90% of the topics in the course, they are considered to have completed the course. This research focuses on predicting the dropout occurrence, so we considered the topics mastered by these 2776 students. The data table holds a student and their entire learning information in one row. Here, we took the entire learning data and separate them by day. The progress (%) in *assessment_report* is the topics the student has mastered concerning the total 383 topics in this College Algebra and Problem-Solving course. The *time_spent* in *timeandtopic_report* is the time spent by the student on that day. We had to group the three datasets by Student ID and spread the information of each day out into multiple columns. This results in a sparse but fine-grained dataset. The complete dataset has 521 attribute columns, with each row holding the whole learning information of a student. The attribute that represents learning progression is the *topics_mastered*. In this complete dataset, we recorded the measurement of this attribute for each day. This attribute is a cumulative score, and Figure 2 (left). shows the rate of learning progression of each student. The changes in the learning progression of a student are the dataset from which we obtained multiple features to be used in the machine learning model. Figure 2 (right) visualizes the rate of changes in learning of each student.

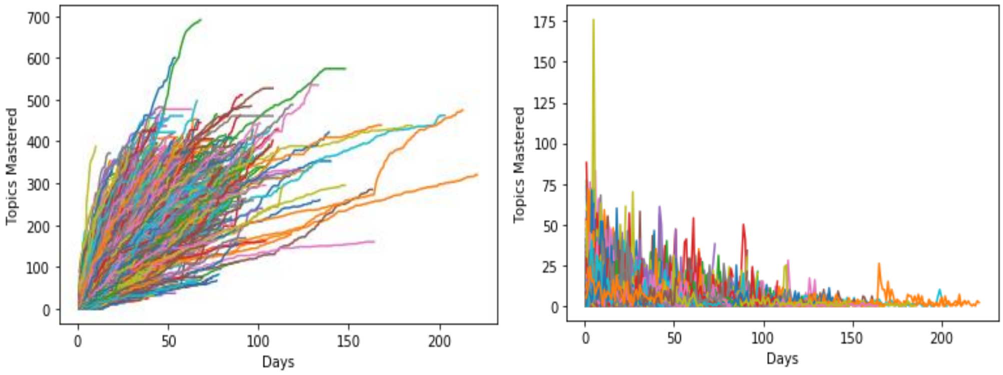

Figure 2. The cumulative rate of learning progression of each student (**left**). The rate of changes in learning progression of each student (**right**).

The predictive modeling employed in this problem is sequence classification [65], depicted in Figure 3. The data needed for this experiment is derived from the learning progression data, which is the *topics_mastered* in *timeandtopic_report*.

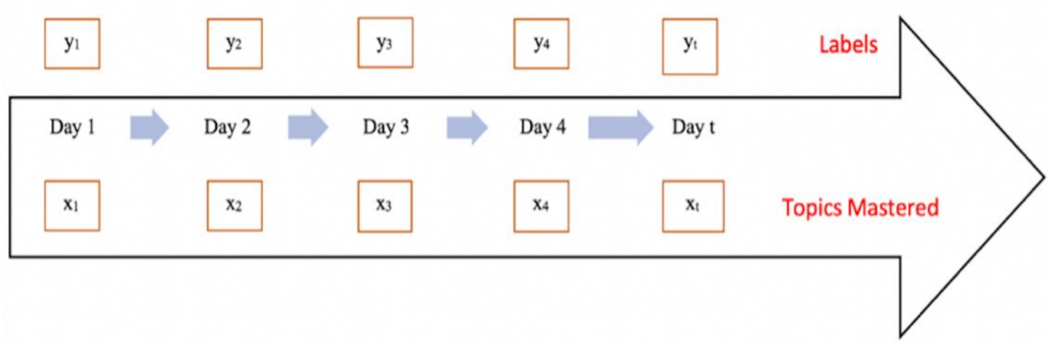

Figure 3. The Data Preprocessing Method: Sequence Classification.

Here, each x is the sequence of topics mastered by the student from the start of the course to the day in consideration, and each y is the label for that sequence, implying if the student continued in the course or dropped out of the course on that day. This gives the preprocessed data needed for the ML model. Most of the attributes of the preprocessed data are of different data types, and so it is cleaned to be of the same numerical data type for modeling [61].

3.2.4. Feature Engineering

The dataset is transformed into a feature table where each row depicts the rate of student learning on any given day from the start of the course. Further, a target column is created in the table, which depicts the label of the learning of the student on that row. The target column was labeled as a "1" if that row of learning led to a student dropout on a particular day and labeled as "0" if that student had a day of learning after that day, which means that the student continued in the course.

From the cleaned data, the topics mastered on each day that the student learns during a course is represented by the array of *topics_mastered*. The rate of student learning on a particular day is represented by a set of statistical features from the array. The most common statistical features employed are the average, standard deviation, variance, skewness, and kurtosis [66,67]. The graph of student learning expressed as topic mastered over time is shown in Figure 4. We can see that the curve is very rough, so average alone cannot aptly represent the rate of learning. Hence, the average calculated in windows through the time series was obtained. This list of averages along with the average can give an overview of the rate of learning and is called the moving average, and the normalized value of this list is used as a feature. Three moving averages with different window sizes, along with the average, were considered. This gives four features to represent the rate of changes in learning. Since the curve in Figure 4 is very rough, four features, namely *skew*, *standard deviation*, *variance*, and *kurtosis*, are used to represent this roughness. The other features used in the analysis are *topics_mastered*, overall trajectory, and final trajectory. The relationship between the first and the last day in distribution is calculated as overall trajectory The relationship between the last two days in distribution is calculated as the final trajectory.

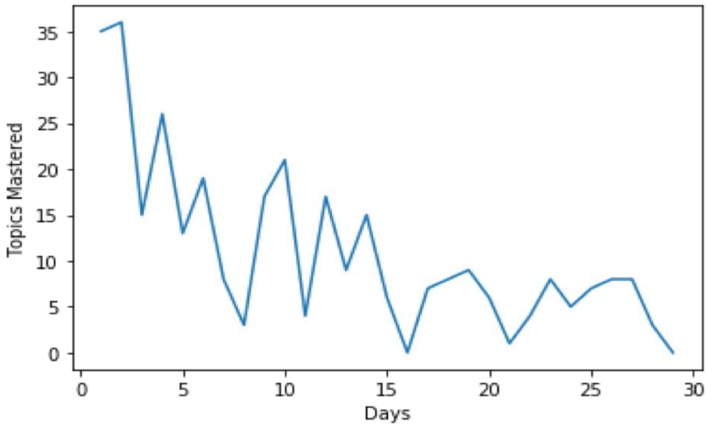

Figure 4. The Graph of Student Learning expressed as topic mastered over time.

3.3. Machine Learning Modeling

Once the feature table is created, it holds the data for the machine learning model. The ML modeling uses the given input features to perform the prediction of dropout of MOOC students. The ML modeling has three steps:

1. Feature selection and model fitting,
2. Prediction model training, and
3. Prediction model testing.

3.3.1. Feature Selection and Model Fitting

In this step, the features generated were evaluated and validated. To predict the student learning outcomes in MOOC, Exploratory Data Analysis (EDA) technique, called the correlation matrix method, was used to validate the features [16]. This in turn would avoid data leakage in the ML model and increase the success of the modeling experiment.

Once the data is balanced, the *sci-kit* learn tool was used to split these vectors into the training features, training labels, testing features, and testing labels. From the generated data, 75% was used for training the model, while the remaining 25% was used to test the model.

3.3.2. Prediction Model Training

The Random Forest (RF) ensemble learning approach was chosen for this method due to its robustness to outlier data and the ability to inspect the model for insights into the most powerful discriminating variables. These aspects support our longer-term research goal to take these insights and target interventions at these variables.

The most used ML models in EDM include XGBoost, RF, and SVM [53]. Among the ML models in EDM, RF performs better [16,53]. The accuracy of the RF model is as good as or sometimes even better than most of the ML models [68,69]. RF is more robust to outliers and noise. The internal estimates of error, strength, correlation, and variable importance are very useful in this research. The classification trees in RF make use of Gini impurity to reduce the errors in prediction by the decision trees. RF works to reduce this value for each tree, thereby reducing overfitting and data bias errors. This makes RF very robust when predicting the noisy dataset with many outliers. The dataset in this research, although it goes through proper feature engineering and feature selection processes, still holds a great deal of outliers, and hence, random forest is better suited for this research. Gini impurity, since attached to each feature, provides individual predictor importance values. RF methodology is highly suitable for use in classification problems when the goal of the methodology is to produce an accurate classifier and to provide insight regarding the discriminative ability of individual variables [70]. This research focuses on the feature engineering approach and its contribution, hence making RF the prime choice for the machine learning model. Here, the RF model was trained with the training data and was ready to predict whether the student will drop out of the course or continue the course when features of learning of a student in MOOC were input to the model.

3.3.3. Prediction Model Testing

Once the model was trained, we performed experiments to test the model for its performance and see if the model could predict the correct outcome for a given set of features. In this study, we performed five experiments to test the performance of the model with different sets of inputs. The experiments were set up in a way to test the model in both the edge case scenarios as well as normal case scenarios. Further, these experiments check whether the model performs expectedly. The details of these five experiments are explained in the Experimental Setup section, and the results from these experiments are described in the Model Evaluation section.

3.4. Model Evaluation

There are three main processes of model evaluation employed in this research. They are model validation, feature importance, and SHAP plot.

3.4.1. Model Validation

Different methods have been reported to validate the models, including direct correlation, Cohens Kappa, and Accuracy [71], but accuracy is not recommended for evaluating the model because it depends on the base rates of different classes [72]. It is important to calculate the missed calculations to measure the sensitivity of the classifier using recall. Moreover, for the evaluation of a prediction model, a combined method, such as F1-score, which considers both true and false classification results based on precision and recall, is the better metric. We also performed the model validation using these four metrics [73].

Even though accuracy alone cannot be used to validate a model, it portrays the performance of the model, therefore the accuracy of the model was also calculated. The Receiver Operating Characteristic (ROC) curve is a plot to show the predictive power of binary classifier models [74]. This curve is obtained by plotting the True-Positive Rate to the False-Positive Rate. With this curve, we can also see the Area Under the Curve. The Area Under the Curve (AUC) is the other validation method employed in evaluating a prediction model. A value of at least 0.7 for these metrics is accepted in the research community.

3.4.2. Feature Importance

To predict retention of students in MOOCs, the feature importance method was used as an iterative process to identify important features for the prediction model RF classifier [75]. The success of this evaluation method motivated its use in the feature selection performed in the research.

This evaluation method is a visualization technique used to analyze the features used in the model. Every model has a coefficient score attached to a feature after its training by calculating the Gini impurity. The feature with the highest coefficient value associated with the model is the most important contributor to the prediction. All *scikit-learn* models generate a coefficient summary, which is used to plot a histogram plot in this research to visualize the importance of the features used. This can be an iterative process, where the more important feature can be selected over the less important feature if there is a dependency established between them.

3.4.3. SHAP Plot

SHAP plots are visualizations used to identify the most important contributor to the model's predictions. SHAP is a relatively new visualization technique used to evaluate the features used in the machine learning model for individual predictions. It plays an important role in visualizing the contribution of features towards the prediction by the model [76]. These plots show the feature as they contribute to either the positive or the negative class in the prediction and how the model is moved step by step by the features towards its predictions.

4. Results and Discussions

4.1. Data Extraction

Among the 3172 students, only 396 students completed this course, while the remaining 2776 students did not due to some reason. This shows that only 12.5% completed the course successfully, and 87.5% of students in this course dropped out.

The data is in 4 different reports:

1. *class_report*,
2. *assessment_report*
3. *progress_report*
4. *timeandtopic_report*

The *class_report* is the highest-level data that was not used for the analysis, while *assessment_report*, *progress_report*, and *timeandtopic_report* were grouped with student ID as the key. The attributes considered from the datasets are tabulated in Table 2.

Table 2. Attributes in Dataset.

Attributes	Description
Student ID	Student primary key
time_and_topics	the time taken and the topics mastered for a day
topics_mastered	the topics mastered for a day
topics_practiced	the topics practiced by the student for a day
time_spent	the time spent by the student for a day

4.2. Feature Engineering

Eleven features were considered to represent the rate of student learning in the MOOC course till the day of consideration for analysis, as shown in Table 3.

Table 3. Features Engineered in this Research.

Average	Kurtosis	Overall Trajectory
Standard Deviation Variance Skew	Moving average with window size 2 Moving average with window size 3 Moving average with window size 4	Final Trajectory Days in consideration

These 11 features were found to make the learning features of a student on a day. A small sample of the final feature table is shown in Table 4.

Table 4. A Sample Feature Table.

Mov Avg 2	Mov Avg 3	Mov Avg 4	Skew	Overall Trajectory	Final Trajectory	Average	Standard Deviation	Variance	Kurtosis	Day
0	0	0	0	0	0	0	0	0	−3	1
0	0	0	0	1.5707	1.5707	0	0	0	−3	2
2	1.3333	0	0.707	1.5707	1.5707	1.3333	1.8856	3.5555	−1.5	3
3.6055	2.4037	1.5	0.493	1.5707	0.4636	1.5	1.6583	2.75	−1.3719	4
5.0249	4.3843	3.1324	0.152	1.5707	1.1902	2.2	2.0396	4.16	−1.6268	5

4.3. Feature Selection and Model Fitting

The correlation between the 11 features is established, as shown in Figure 5. Three groups of features are very dependent on each other. The three groups of features are (1) moving averages; (2) the average, standard deviation, and variance; and (3) kurtosis and skew. The dependency value between kurtosis and day features falls in the range of 0.8 and above. Hence, to remove this dependency, a trial run on an RF ML model was run, and the feature importance plot for this set of features was obtained and shown in Figure 5.

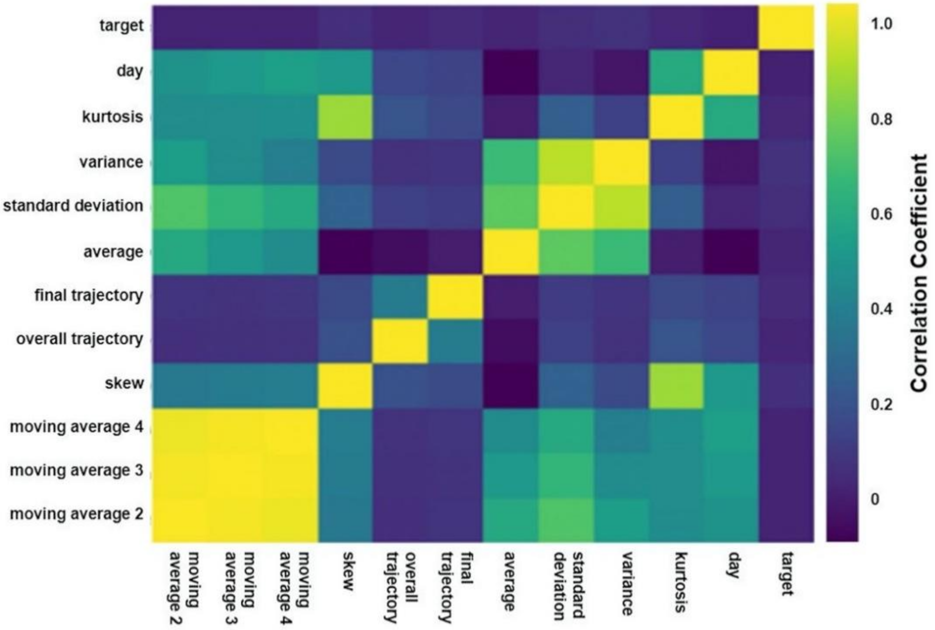

Figure 5. Correlation Matrix of Features.

From Figure 6, the most important feature in each of the three dependent feature groups was selected. The features moving average with window size 2, skew, and average was selected, and other features were removed from the group. Once again, the correlation between the features after the feature selection was tried, and the correlation matrix obtained after feature selection is shown in Figure 7.

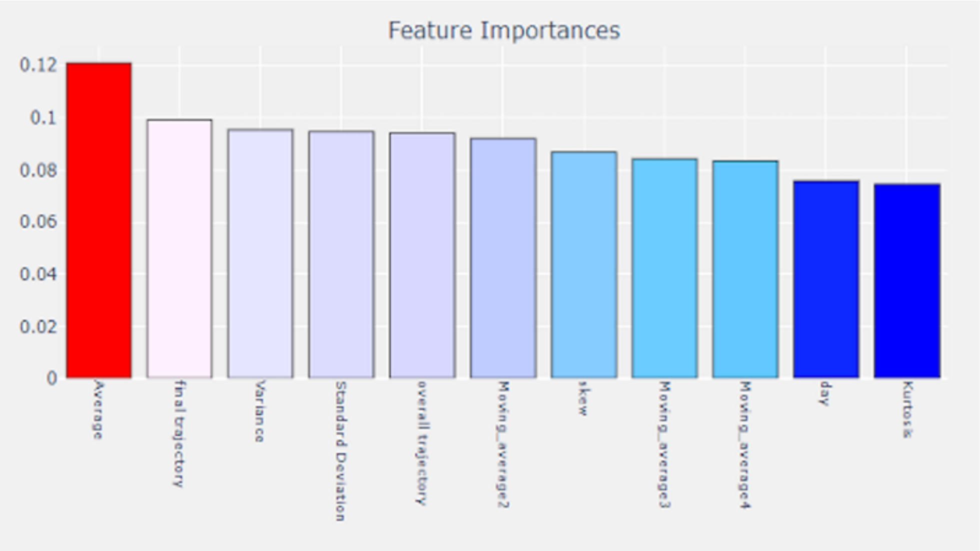

Figure 6. Feature Importance Plot.

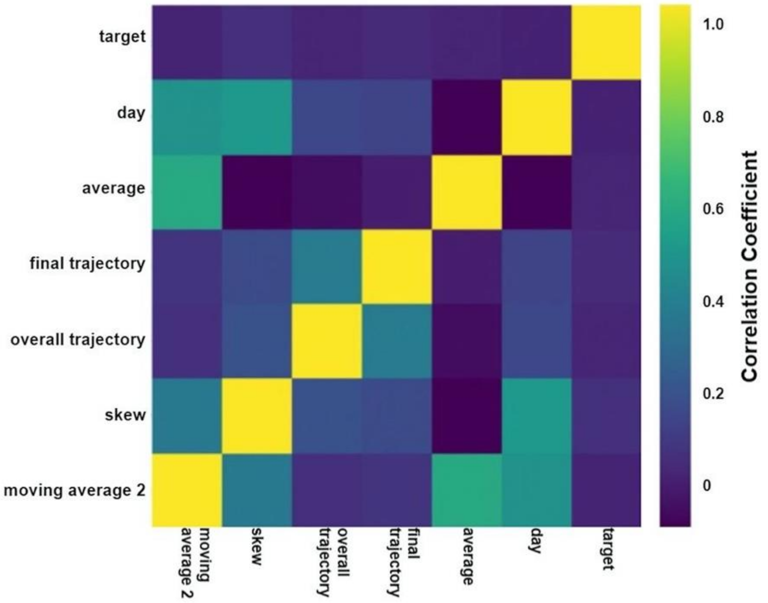

Figure 7. Correlation Matrix of Features after Feature Selection.

The target spread is shown in Table 5. From the correlation matrix obtained after the feature selection, there is no correlation between two features with more than 0.5, and all the features are completely independent of the target variables.

Table 5. Target Values.

Target Value	Num Data Points
0	39,529
1	2776

The *sci-kit* learn tool was now used to split these vectors into the training features, training labels, testing features, and testing labels. From this, 75% of the data was used for training the model, and the remaining 25% of the data was used to test the model.

Here, "0" represents the continue label, while "1" represents the dropout label, as explained in the feature engineering section. The table shows an imbalance in data points where 93.5% of the data points are for continue, while 6.5% of the data points are for dropout. In this research, we also employed SMOTE to overcome this imbalance [77,78]. The balance in the data after the application of SMOTE is shown in Table 6.

Table 6. Target values after SMOTE.

Target Value	Number of Data Points
0	39,529
1	39,529

4.4. Model Training

The RF model was trained from scikit-learn with the specifications shown in Table 7.

Table 7. Random Forest Model Specifications.

Arguments	Value	Specification
n_estimators	1000	Number of trees
max_features	auto	sqrt (number of features)
random_state	42	Control the randomness
criterion	Gini	Gini impurity

This research uses 1000 decision trees, as the number of trees was directly proportional to the model performance, and the time taken to train more than 1000 trees is too long that it becomes impractical in application. The random state is set to 42 so that when the randomness is fixed, the research can be replicated with the same results, and this is the random state used throughout the experiment. The Gini impurity criterion was used to obtain the feature importance plot, as shown in Figure 6 in the above section. The maximum number of features that the model considers is set as "auto", which is a fixed value of the square root of the number of features used in the model. This is also fixed to be able to reproduce the results obtained from this experiment. Once the model was trained with these specifications shown in Table 5, model testing was performed.

4.5. Model Testing

The model was tested with the testing data to validate the overall performance of the model. The spread of the test data point for this experiment can be seen in Table 8.

Table 8. Testing Data Point Spread.

Target Value	Number of Data Points
0	9883
1	9882

The results show the overall performance of the model as well as the feasibility of this model to predict the dropout of a student.

4.6. Model Validation

This experiment involved testing the model with the entire testing data separated from the training data before training the model. We passed the input testing data to the model and received its predictions for this set of input. Then, we checked the output testing data values with the model's prediction values and used them to calculate the model validation metrics. The results of the model validation are shown in Table 9.

Table 9. The Results of the Model Validation.

Class	Precision	Recall	F1-Score	Support
0	0.91	0.84	0.87	9883
1	0.85	0.91	0.88	9882

The accuracy and AUC of the prediction model are:

- ➢ Accuracy = 0.87665
- ➢ AUC = 0.94654

The AUC is plotted along with the line representing the True-Positive Rate of 0.5 and the False-Positive Rate of 0.5 to show the performance of the model, and this method of validation is called the ROC curve analysis. Figure 8 shows the result of the ROC curve analysis performed for the model trained and tested in this research, and the AUC is far away from the 0.5 line, which means that the model covered the dataset well and can predict the student dropout or continue for most cases in the dataset.

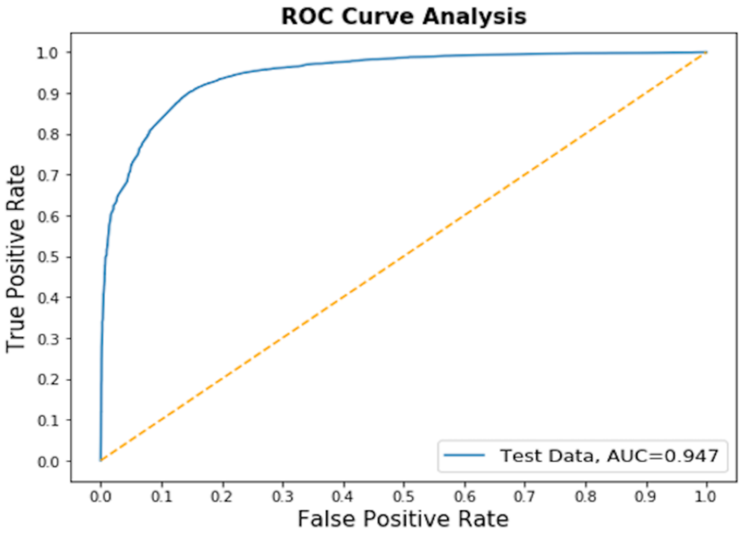

Figure 8. The ROC of the Model.

The variation of accuracy, precision, recall, and F1-score of the model for different days are shown in Figure 9. It can be observed that the accuracy of the model is consistently above 70% and mostly above 80%, the precision of the model is always above 70% and consistently above 80% and mostly above 90%, the recall of the model is always above 80% and consistently above 90%, and the F1-score of the model is always above 70% and consistently above 80% and mostly above 90%, respectively.

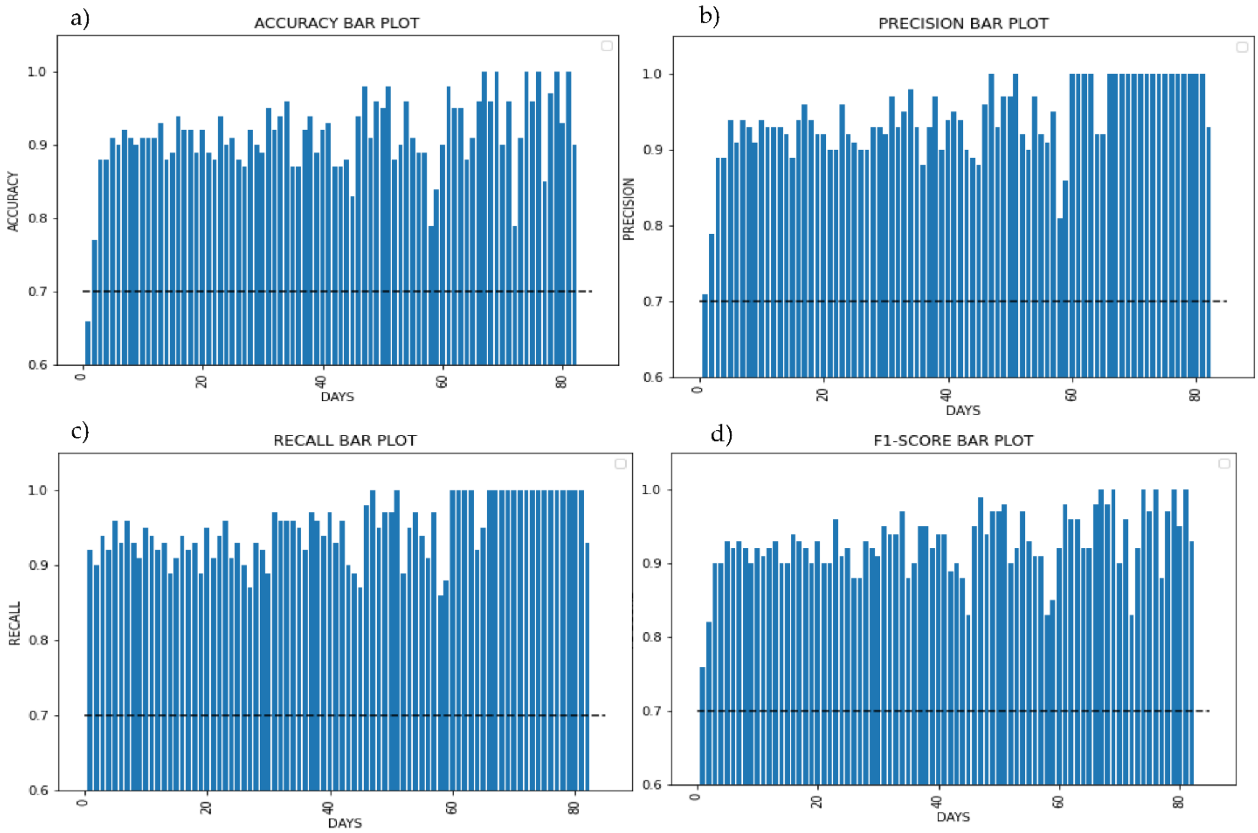

Figure 9. (**a**). Accuracy of the Model on Different Days (**b**). Precision of the Model on Different Days (**c**). Recall of the Model on Different Days (**d**). F1-score of the Model on Different Days.

These results show that the model performs well for any given set of data, as the dataset has less data as the number of days increases, but this is not reflected on the performance of the model, showing the robustness of the model. However, even with these results, the model cannot be explained. Hence, this research uses the SHAP visualizations to explain the random forest model trained and tested in this research.

4.7. SHAP Visualizations

This research uses the SHAP python library to visualize the impact of the features used in the prediction model. When the trained model is given to the SHAP library with the testing input features, it gives the following Figure 10. From Figure 10, we can obtain the following inferences:

➢ High values of average topics mastered by the students point towards a continuation of the course, while low values of average topics mastered by the students point towards dropout from the course.

➢ High values of final trajectory in topics mastered by the students point towards a continuation of the course, while low values of final trajectory in topics mastered by the students point towards dropout from the course.

➢ Low values of skew in topics mastered by the students point towards a continuation of the course, while high values of skew in topics mastered by the students point towards dropout from the course.

➢ Low values of moving average of window size 2 in topics mastered by the students point towards a continuation of the course, while high values of moving average of window size 2 in topics mastered by the students point towards dropout from the course.

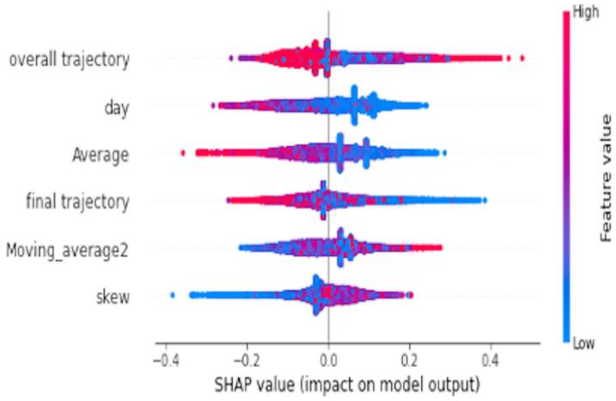

Figure 10. The SHAP Summary Plot for this Prediction Model.

To better understand the feature interactions, we present the SHAP force plots for two different data point examples, shown below. It can help visualize how these features interact with each other while the model arrives at its prediction. We input the data as shown in Table 10.

Table 10. Input Values for SHAP Force Plot 1.

Features	Values
moving_average 2	0.7675
skew	0.7071
overall trajectory	0
final trajectory	1.5707
average	0.5117
day	3

For the input in the above table, the model correctly predicted dropout, class 1. The SHAP force plot shown in Figure 11 helps visualize the feature interactions that lead to this prediction.

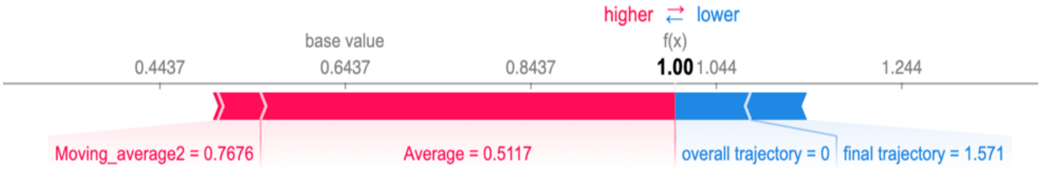

Figure 11. SHAP Force plot 1.

In Figure 11, we can see the moving_average 2, average, and skew pushing the model to predict 1, while the overall trajectory and final trajectory push the model to predict 0. Because the features are aptly weighted, the model arrives at its correct prediction.

Consider another set of inputs, shown in Table 11.

Table 11. Input Values for SHAP Force Plot 2.

Features	Values
moving_average 2	35.2411
skew	0.3551
overall trajectory	0.0182
final trajectory	1.4272
average	5.7054
day	30

For the input in the above table, the model correctly predicts dropout, class 1. The SHAP force plot shown in Figure 12 helps visualizes the feature interactions that lead to this prediction.

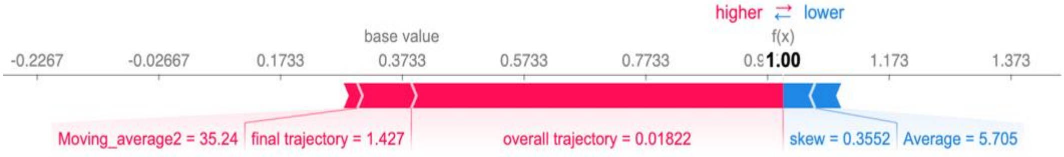

Figure 12. SHAP Force Plot 2.

In Figure 12, we observe the moving_average 2, final trajectory, and overall trajectory pushing the model to predict 1, while skew and average push the model to predict 0.

5. Discussion

The results from the model validation and testing show that the model can predict the student dropout accurately given the feature set of any rate of changes in the student learning. These results answer the research questions put forward by this research:

> *RQ 1: What are the features of changes in learning progression that are associated with students who drop out of a MOOC course?*

The SHAP visualizations and the feature importance from RF point out the features and their impact on the prediction made by the ML model. The results show that lower average values, lower final trajectory values, higher skew values, and higher values of moving average with a window size of two days are features that are associated with student dropouts, and these features help us in predicting this occurrence.

> *RQ 2: Given a set of features of changes in the learning progression of a student on a day of consideration, can we predict the day of dropout of a student in a MOOC course?*

The results from validating the model show that it is possible to predict the student dropout given a set of features of changes in student learning. The results from validating the model show that it is possible to predict the student dropout with an accuracy of 87.6% given the set of features of changes in the student learning used in this research, and this is comparable with the previously reported accuracies of 81.8% [53], 86.5% [79], and 87.6% [68]. This shows that if the learning progression data are available, the features of changes in the student learning should be considered to predict the student dropout. It is observed from the SHAP visualizations that lower average values, lower final trajectory values, higher skew values, and higher values of moving average with a window size of two days are traits of a student on the day they drop out when compared to the days the student continues in the course.

A key insight of this research is the use of learner progression through frequent micro-assessments. While many studies use time-oriented features, these are usually tied to learner engagement through clickstream data, as described in the Related Work section. This work considers the trajectory, or pattern of student learning. The closest similar work we have seen is from [80], who examine the greater impact of short-term "near in time" events compared to long-term events on dropout forecasts using a time-controlled Long-Short-Term Memory neural network (E-LSTM). The authors claimed a 97% accuracy result of this model, which is a strong result. However, this result is still based on engagement data and not assessment data. While it is useful to know if learner behaviors can be monitored to predict dropout, it is hardly surprising that greater engagement leads to greater success. In addition, learner behavior models have an implicit closed-world assumption when it comes to learner engagement—namely that the system can track all the engagement events. Yet, we know this is not possible; it does not account for self-study outside of the MOOC platform or for learning styles of students that are less outgoing and therefore do not participate in online discussion forums and the like. Our approach focuses on progression data, with the recency of successful completion of learning gates (the moving average) as a strong predictor of continued success (and no dropout).

6. Conclusions

Learning analytics has earned considerable attention in EDM and in particular the prediction of student dropout using ML application. Although learner enrollment in MOOCs has been increasing progressively, low completion rates remain a major problem. The prediction of the dropout of learners will help educational administrators evaluate and comprehend the learning activities of learners through the different interactions of the learners. It will also enable educational administrators to develop approaches to promote and deliver learner remediation. The results from this research demonstrate that we can provide reliable prediction based on six easily obtainable features via an ML approach using RF for prediction, which this research hopes could be easily and reliably implemented across various courses from different domains. As discussed in the results section, this research is successful in predicting the student dropout from MOOC given the set of features used in this research, and to the best of our knowledge, there is no such stable and accurate predictive methodology.

The dataset used in this research is derived from the self-paced math course College Algebra and Problem Solving offered on the MOOC platform Open edX offered by Arizona State University (ASU). It consists of students taking this course starting from March 2016 to March 2020. The dataset is analyzed using RF; the feature and modeling evaluation is done by Precision, Recall, F1-score, AUC, and ROC curve; and the model is explained by SHAP. This model can predict the student dropout at an acceptable standard in the research community with an accuracy of 87.6%, precision of 85%, recall of 91% and F1-score of 88%, and an AUC of 94.6%.

This work, like the works discussed in the Related Work section, focuses on machine learning approaches to predicting MOOC dropout and success. As Ahmed et al. [81] recently pointed out in their reflections on the last decade of the plethora of MOOC research, few MOOCs employ formative feedback during the learning progression to improve effort and achievement. Machine learning models are only useful if applied in context to encourage higher retention and success rates. For future work, in addition to continued refinement of this model and potentially generalizing beyond the STEM course application we have developed this model on, we are also interested in using the model to design interventions. The power of a model based on learner progression is that it provides key insights into *when* a learner may be at risk of dropping out, so a just-in-time (JIT) intervention may be designed to improve retention and success. We believe powerful Learning Analytics models coupled with causal approaches, such as that of [82], will result in specific, targeted JIT interventions personalized to the context of individual learners.

Author Contributions: Conceptualization, S.D. and K.G.; Data curation, S.D.; Formal analysis, S.D.; Investigation, S.D.; Methodology, S.D. and J.C.; Project administration, K.G. and J.C.; Resources, J.C.; Software, S.D.; Supervision, K.G.; Validation, J.C.; Visualization, S.D.; Writing—original draft, S.D. and K.G.; Writing—review & editing, K.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The work in this study is covered under ASU Knowledge Enterprise Development IRB titled *Learner Effects in ALEKS*, STUDY00007974.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from EdPlus and are available from the authors with the permission of EdPlus.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Distribution of Students Across Different Age Groups.

Ranges of Ages	Number of Students	Success	Dropout
0–9	1	0	1
10–19	364	101	263
20–29	1703	147	1556
30–39	737	50	687
40–49	231	14	217
50–59	91	7	84
60–69	20	3	18
≥70	0	0	0

Table A2. Distribution of Students Across Different Gender Groups.

Gender	Number of Students	Success	Dropout
Female	1502	102	1400
Male	1204	138	1066

Table A3. Distribution of Students Across Different Ethnic Groups.

Ethnicity	Number of Students	Success	Dropout
White	1155	114	1041
Black	335	17	318
Hispanic, White	241	27	214
Hispanic	237	18	219
Asian	131	21	110
Black, White	41	3	38
Black, Hispanic	23	0	23
American I	20	1	19
Asian, White	18	3	15
American I, White	13	1	12
Asian, Black	11	1	10
American I, Hispanic	11	0	11
Black, Hispanic, White	10	1	9
Haw/Pac	10	0	10
American I, Hispanic, White	8	0	8
Asian, Haw/Pac	6	0	6
Haw/Pac, Hispanic	6	0	6
Asian, Hispanic, White	5	0	5
Asian, Hispanic	5	2	3
Haw/Pac, White	4	0	4
American I, Black	3	0	3
Asian, Haw/Pac, White	3	0	3
Asian, Haw/Pac, Hispanic	2	0	2
American I, Black, White	2	0	2
Asian, Black, Haw/Pac, Hispanic	1	0	1
American I, Black, Hispanic	1	0	1
American I, Asian, Black, White	1	0	1
Asian, Black, Hispanic	1	0	1
Black, Haw/Pac	1	0	1
Haw/Pac, Hispanic, White	1	0	1

References

1. Rolfe, V. A Systematic Review of The Socio-Ethical Aspects of Massive Online Open Courses. *Eur. J. Open Distance E-Learn.* **2015**, *18*, 52–71. [CrossRef]
2. Kumar, J.A.; Al-Samarraie, H. An Investigation of Novice Pre-University Students' Views towards MOOCs: The Case of Malaysia. *Ref. Libr.* **2019**, *60*, 134–147.
3. Nagrecha, S.; Dillon, J.Z.; Chawla, N.V. MOOC dropout prediction: Lessons learned from making pipelines interpretable. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; pp. 351–359. [CrossRef]
4. Qiu, J.; Tang, J.; Liu, T.X.; Gong, J.; Zhang, C.; Zhang, Q.; Xue, Y. Modeling and Predicting Learning Behavior in MOOCs. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, 22–25 February 2016; pp. 93–102. [CrossRef]

5. Dalipi, F.; Imran, A.S.; Kastrati, Z. MOOC dropout prediction using machine learning techniques: Review and research challenges. In Proceedings of the IEEE Global Engineering Education Conference, Santa Cruz de Tenerife, Spain, 17–20 April 2018; pp. 1007–1014. [CrossRef]
6. Kim, T.-D.; Yang, M.-Y.; Bae, J.; Min, B.-A.; Lee, I.; Kim, J. Escape from infinite freedom: Effects of constraining user freedom on the prevention of dropout in an online learning context. *Comput. Hum. Behav.* **2017**, *66*, 217–231. [CrossRef]
7. Shah, D. By the Numbers: MOOCS in 2018 Class Central. 2018. Available online: https://www.classcentral.com/report/mooc-stats-2018/ (accessed on 16 December 2018).
8. Feng, W.; Tang, J.; Liu, T.X. Understanding Dropouts in MOOCs. In Proceedings of the 23rd American Association for Artificial Intelligence National Conference (AAAI), Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 517–524. [CrossRef]
9. Hellas, A.; Ihantola, P.; Petersen, A.; Ajanovski, V.V.; Gutica, M.; Hynninen, T.; Knutas, A.; Leinonen, J.; Messom, C.; Liao, S.N. Predicting academic performance: A systematic literature review. In Proceedings of the Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus, 2–4 July 2018; pp. 175–199. [CrossRef]
10. Baker, R.S.; Yacef, K. The state of educational data mining in 2009: A review and future visions. *J. Educ. Data Min.* **2009**, *1*, 3–17.
11. West, D.M. Big Data for Education: Data Mining, Data Analytics, and Web Dashboards. *Gov. Stud. Brook.* **2012**, *4*, 1–10.
12. Lara, J.A.; Lizcano, D.; Martínez, M.A.; Pazos, J.; Riera, T. A system for knowledge discovery in e-learning environments within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA. *Comput. Educ.* **2014**, *72*, 23–36. [CrossRef]
13. Chakraborty, B.; Chakma, K.; Mukherjee, A. A density-based clustering algorithm and experiments on student dataset with noises using Rough set theory. In Proceedings of the IEEE International Conference on Engineering and Technology, Coimbatore, India, 17–18 March 2016; pp. 431–436. [CrossRef]
14. Chauhan, N.; Shah, K.; Karn, D.; Dalal, J. Prediction of student's performance using machine learning. In Proceedings of the 2nd International Conference on Advances in Science & Technology, Mumbai, India, 8–9 April 2019.
15. Salloum, S.A.; Alshurideh, M.; Elnagar, A.; Shaalan, K. Mining in Educational Data: Review and Future Directions. In Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), Cairo, Egypt, 8–10 April 2020.
16. Al-Shabandar, R.; Hussain, A.; Laws, A.; Keight, R.; Lunn, J.; Radi, N. Machine learning approaches to predict learning outcomes in Massive open online courses. In Proceedings of the International Joint Conference on Neural Networks, Anchorage, AK, USA, 14–19 May 2017; pp. 713–720. [CrossRef]
17. Baker, R.S.J.D.; Siemens, G. Educational Data Mining and Learning Analytics. In *Cambridge Handbook of the Learning Sciences*, 2nd ed.; Keith Sawyer, R., Ed.; Cambridge University Press: New York, NY, USA, 2014; pp. 253–274.
18. Fiaidhi, J. The Next Step for Learning Analytics. *IT Prof.* **2014**, *16*, 4–8. [CrossRef]
19. Gašević, D.; Rose, C.; Siemens, G.; Wolff, A.; Zdrahal, Z. Learning Analytics and Machine Learning. In Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, Indianapolis, IN, USA, 24–28 March 2014; pp. 287–288. [CrossRef]
20. Liyanagunawardena, T.R.; Parslow, P.; Williams, S. Dropout: MOOC participant's perspective. In Proceedings of the EMOOCs 2014, the Second MOOC European Stakeholders Summit, Lausanne, Switzerland, 10–12 February 2014; pp. 95–100.
21. Jayaprakash, S.M.; Moody, E.W.; Lauría, E.J.M.; Regan, J.R.; Baron, J.D. Early alert of academically at-risk students: An open source analytics initiative. *J. Learn. Anal.* **2014**, *1*, 6–47. [CrossRef]
22. Márquez-Vera, C.; Cano, A.; Romero, C.; Noaman, A.Y.M.; Fardoun, H.M.; Ventura, S. Early dropout prediction using data mining: A case study with high school students. *Expert Syst.* **2016**, *33*, 107–124. [CrossRef]
23. Palmer, S. Modelling engineering student academic performance using academic analytics. *Int. J. Eng. Educ.* **2013**, *29*, 132–138.
24. Papamitsiou, Z.; Economides, A. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educ. Technol. Soc.* **2014**, *17*, 49–64.
25. Peña-Ayala, A. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Syst. Appl.* **2014**, *41*, 1432–1462. [CrossRef]
26. Zacharis, N.Z. A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *Internet High. Educ.* **2015**, *27*, 44–53. [CrossRef]
27. Cen, L.; Ruta, D.; Powell, L.; Hirsch, B.; Ng, J. Quantitative approach to collaborative learning: Performance prediction, individual assessment, and group composition. *Int. J. Comput. Collab. Learn.* **2016**, *11*, 187–225. [CrossRef]
28. Mueen, A.; Zafar, B.; Manzoor, U. Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *Int. J. Mod. Educ. Comput. Sci.* **2016**, *8*, 36. [CrossRef]
29. Huang, S.; Fang, N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Comput. Educ.* **2013**, *61*, 133–145. [CrossRef]
30. Marbouti, F.; Diefes-Dux, H.; Madhavan, K. Models for early prediction of at-risk students in a course using standards-based grading. *Comput. Educ.* **2016**, *103*, 1–15. [CrossRef]
31. Tempelaar, D.; Rienties, B.; Giesbers, B. In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Comput. Hum. Behav.* **2015**, *47*, 157–167. [CrossRef]
32. Kizilcec, R.F.; Pérez-Sanagustín, M.; Maldonado, J.J. Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Comput. Educ.* **2017**, *104*, 18–33. [CrossRef]

33. Kuzilek, J.; Hlosta, M.; Herrmannova, D.; Zdrahal, Z.; Wolff, A. Ou analyse: Analysing at-risk students at the open university. *Learn. Anal. Rev.* **2015**, *8*, 1–16.
34. Wolff, A.; Zdrahal, Z.; Nikolov, A.; Pantucek, M. Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In Proceedings of the Third Conference on Learning Analytics and Knowledge, Leuven, Belgium, 8–12 April 2013.
35. Hlosta, M.; Herrmannova, D.; Vachova, L.; Kuzilek, J.; Zdrahal, Z.; Wolff, A. Modelling student online behaviour in a virtual learning environment. *arXiv Prepr.* **2018**, arXiv:1811.06369.
36. Cui, Y.; Chen, F.; Shiri, A. Scale up predictive models for early detection of at-risk students: A feasibility study. *Inf. Learn. Sci.* **2020**, *121*, 97–116. [CrossRef]
37. Soffer, T.; Cohen, A. Student's engagement characteristics predict success and completion of online courses. *J. Comput. Assist. Learn.* **2019**, *35*, 378–389. [CrossRef]
38. Winne, P. Improving Measurements of Self-Regulated Learning. *Educ. Psychol.* **2010**, *45*, 267–276. [CrossRef]
39. Sha, L.; Looi, C.-K.; Chen, W.; Zhang, B. Understanding mobile learning from the perspective of self-regulated learning. *J. Comput. Assist. Learn.* **2012**, *28*, 366–378. [CrossRef]
40. Tang, J.; Xie, H.; Wong, T. *A Big Data Framework for Early Identification of Dropout Students in MOOC. Technology in Education. Technology-Mediated Proactive Learning*; Springer: Berlin/Heidelberg, Germany, 2015.
41. Amnueypornsakul, B.; Bhat, S.; Chinprutthiwong, P. Predicting attrition along the way: The UIUC model. In Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, Dota, Qatar, 25–29 October 2014; pp. 55–59.
42. Baker, R.; Evans, B.; Li, Q.; Cung, B. Does Inducing Students to Schedule Lecture Watching in Online Classes Improve Their Academic Performance? An Experimental Analysis of a Time Management Intervention. *Res. High. Educ.* **2018**, *60*, 521–552. [CrossRef]
43. Cicchinelli, A.; Veas, E.; Pardo, A.; Pammer-Schindler, V.; Fessl, A.; Barreiros, C.; Lindstädt, S. Finding traces of self-regulated learning in activity streams. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge, Sydney, NSW, Australia, 7–9 March 2018; pp. 191–200.
44. Lim, J.M. Predicting successful completion using student delay indicators in undergraduate self-paced online courses. *Distance Educ.* **2016**, *37*, 317–332. [CrossRef]
45. Park, J.; Denaro, K.; Rodriguez, F.; Smyth, P.; Warschauer, M. Detecting changes in student behavior from clickstream data. In Proceedings of the 7th International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, 21–30 March 2017. [CrossRef]
46. Gašević, D.; Dawson, S.; Rogers, T.; Gasevic, D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet High. Educ.* **2016**, *28*, 68–84. [CrossRef]
47. Bozkurt, A.; Yazıcı, M.; Aydin, I.E. *Cultural diversity and its implications in online networked learning spaces In Research Anthology on Developing Effective Online Learning Courses*; Information Resources Management Association, Ed.; IGI Global: Hershey, PA, USA, 2018; pp. 56–81.
48. Baker, R.S.; Inventado, P.S. *Educational Data Mining and Learning Analytics in Learning Analytics*; Springer: New York, NY, USA, 2014; pp. 61–75.
49. Kőrösi, G.; Farkas, R. Mooc performance prediction by deep learning from raw clickstream data. In Proceedings of the in International Conference in Advances in Computing and Data Sciences, Valletta, Malta, 24–25 April 2020; pp. 474–485.
50. Hung, J.L.; Wang, M.C.; Wang, S.; Abdelrasoul, M.; Li, Y.; He, W. Identifying at-risk students for early interventions—A time-series clustering approach. *IEEE Trans. Emerg. Top. Comput.* **2015**, *5*, 45–55. [CrossRef]
51. Akçapınar, G.; Hasnine, M.N.; Majumdar, R.; Flanagan, B.; Ogata, H. Developing an early-warning system for spotting at-risk students by using eBook interaction logs. *Smart Learning Environments* **2019**, *6*, 4. [CrossRef]
52. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
53. Adnan, M.; Habib, A.; Ashraf, J.; Mussadiq, S.; Raza, A.A.; Abid, M.; Bashir, M.; Khan, S.U. Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access* **2021**, *9*, 7519–7539. [CrossRef]
54. Costa, E.B.; Fonseca, B.; Santana, M.A.; De Araújo, F.F.; Rego, J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Comput. Hum. Behav.* **2017**, *73*, 247–256. [CrossRef]
55. Cano, A.; Leonard, J.D. Interpretable Multiview Early Warning System Adapted to Underrepresented Student Populations. *IEEE Trans. Learn. Technol.* **2019**, *12*, 198–211. [CrossRef]
56. Burgos, C.; Campanario, M.L.; De La Peña, D.; Lara, J.A.; Lizcano, D.; Martínez, M.A. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Comput. Electr. Eng.* **2018**, *66*, 541–556. [CrossRef]
57. Gupta, S.; Sabitha, A.S. Deciphering the attributes of student retention in massive open online courses using data mining techniques. *Educ. Inf. Technol.* **2018**, *24*, 1973–1994. [CrossRef]
58. Praveena, M.; Jaiganesh, V. A Literature Review on Supervised Machine Learning Algorithms and Boosting Process. *Int. J. Comput. Appl.* **2017**, *169*, 32–35. [CrossRef]
59. Eranki, K.L.; Moudgalya, K.M. Evaluation of web based behavioral interventions using spoken tutorials. In Proceedings of the 2012 IEEE Fourth International Conference on Technology for Education, Hyderabad, India, 18–20 July 2012; pp. 38–45.

60. Kanungo, T.; Mount, D.; Netanyahu, N.; Piatko, C.; Silverman, R.; Wu, A. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892. [CrossRef]
61. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* **1996**, *39*, 27–34. [CrossRef]
62. Saa, A.A. Educational Data Mining & Students' Performance Prediction. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 212–220.
63. Canfield, W. ALEKS: A Web-based intelligent tutoring system. *Math. Comput. Educ.* **2001**, *35*, 152.
64. Craig, S.D.; Hu, X.; Graesser, A.C.; Bargagliotti, A.E.; Sterbinsky, A.; Cheney, K.R.; Okwumabua, T. The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors. *Comput. Educ.* **2013**, *68*, 495–504. [CrossRef]
65. Fei, M.; Yeung, D.Y. models for predicting student dropout in massive open online courses. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop, Atlantic City, NJ, USA, 14–17 November 2015; pp. 256–263. [CrossRef]
66. Nanopoulos, A.; Alcock, R.; Manolopoulos, Y. Feature-based classification of time-series data. *Int. J. Comput. Res.* **2001**, *10*, 49–61.
67. Doanne, D.; Seward, L.E. Measuring skewness. *J. Stat.* **2011**, *19*, 1–18.
68. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
69. Alamri, A.; Alshehri, M.; Cristea, A.; Pereira, F.D.; Oliveira, E.; Shi, L.; Stewart, C. *Predicting MOOCs Dropout Using Only Two Easily Obtainable Features from the First Week's Activities. Intelligent Tutoring Systems*; Coy, A., Hayashi, Y., Chang, M., Eds.; Springer: Cham, Switzerland, 2019; pp. 163–173. [CrossRef]
70. Archer, K.J.; Kimes, R.V. Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.* **2008**, *52*, 2249–2260. [CrossRef]
71. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
72. Algarni, A. Data Mining in Education. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 456–461. [CrossRef]
73. Goutte, C.; Gaussier, E. A Probabilistic Interpretation of Precision Recall and F-Score, with Implication for Evaluation. In *Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer Science, 3408*; Losada, D.E., Fernández-Luna, J.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 345–359. [CrossRef]
74. Fawcett, T. Introduction to receiver operator curves. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
75. Sharkey, M.; Sanders, R. A process for predicting MOOC attrition. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 50–54.
76. Bulathwela, S.; Pérez-Ortiz, M.; Lipani, A.; Yilmaz, E.; Shawe-Taylor, J. Predicting Engagement in Video Lectures. In Proceedings of the International Conference on Educational Data Mining, Ifrain, Morocco, 10–13 July 2020.
77. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
78. Wang, J.; Xu, M.; Wang, H.; Zhang, J. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. In Proceedings of the 2006 8th International Conference on Signal Processing, Guilin, China, 16–20 November 2006; Volume 3.
79. Hong, B.; Wei, Z.; Yang, Y. Discovering learning behavior patterns to predict dropout in MOOC. In Proceedings of the 12th International Conference on Computer Science and Education, Houston, TX, USA, 22–25 August 2017; pp. 700–704. [CrossRef]
80. Wang, L.; Wang, H. Learning behavior analysis and dropout rate prediction based on MOOCs data. In Proceedings of the 2019 10th International Conference on Information Technology in Medicine and Education (ITME), Quingdao, China, 23–25 August 2019; pp. 419–423.
81. Yousef, A.M.F.; Sumner, T. Reflections on the last decade of MOOC research. *Comput. Appl. Eng. Educ.* **2021**, *29*, 648–665. [CrossRef]
82. Aldowah, H.; Al-Samarraie, H.; Alzahrani, A.I.; Alalwan, N. Factors affecting student dropout in MOOCs: A cause and effect decision-making model. *J. Comput. High. Educ.* **2019**, *32*, 429–454. [CrossRef]