Predicting Student Dropout in Self-Paced MOOC Course

by

Sheran Dass Dominic Ravichandran

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2021 by the
Graduate Supervisory Committee:

Kevin Gary, Chair
Ajay Bansal
James Cunningham
Adrian Sannier

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

One persisting problem in Massive Open Online Courses (MOOCs) is the issue of student dropout from these courses. The prediction of student dropout from MOOC courses can identify the factors responsible for such an event and it can further initiate intervention before such an event to increase student success in MOOC. There are different approaches and various features available for the prediction of student's dropout in MOOC courses.

In this research, the data 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) from 2016 to 2020 was considered. This research aims to predict the dropout of students from a MOOC course given a set of features engineered from the learning of students in a day. Machine Learning (ML) model used is Random Forest (RF) and this model is evaluated using the validation metrics like accuracy, precision, recall, F1-score, Area Under the Curve (AUC), Receiver Operating Characteristic (ROC) curve. The average rate of student learning progress was found to have more impact than other features. 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. The features engineered in this research are predictive of student dropout and could be used for similar courses to predict student dropout from the course. This model can also help in making interventions at a critical time to help students succeed in this MOOC course.

DEDICATION

To my parents, for all their love and support

ACKNOWLEDGMENTS

I would like to express my gratitude to my thesis mentor Dr. Kevin Gary for his guidance, support, and encouragement throughout my thesis. I am thankful to Dr. Ajay Bansal for being supportive of my research and for being willing to serve on my thesis committee as a member.

I would like to thank Dr. James Cunningham, for providing me with valuable data and his guidance throughout my thesis. Without this dataset, generating and working on this unique problem wouldn't be possible. I would also like to thank Dr. Adrian Sannier for sharing his knowledge and supporting my research.

I would like to thank Julie Greenwood and the leadership of the Action Lab at EdPlus at Arizona State University for their support and access to the ALEKS data used in this study.

I would also like to thank Suddhasvatta Das, a doctoral student under Professor Gary, for all the help during my research and Malavika, my friend, for cheering me on.

Finally, I would like to thank Arizona State University for providing me with amenities as well as permitting me to work with their data which was crucial in the successful completion of my thesis, and the department of CIDSE for the constant guidance and assistance throughout my master's program and thesis.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The Massive Open Online Courses (MOOCs) are 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 (Rolfe, 2015) Further, it has gained importance owing to their flexibility (Kumar & Al-Samarraie 2019) and world-class educational resources (Nagrecha et al. 2017). ASUx®, Coursera®, and Khan Academy® are some examples of MOOC providers. Since 2012, MOOC modalities have received prevalent usage by top Universities (Qiu et al., 2016). Investigations undertaken by such institutions indicated 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 (Qiu et al., 2016).

However, despite the potential benefits of MOOCs, the rate of students who drop out of courses has been typically very high (Dalipi et al., 2018; Kim et al. 2017; Shah, 2018). Recent reports also show that the completion rate in MOOCs is very low as compared to the number of those enrolled in these courses (Feng et al. 2019), Hence the Prediction of the dropout of students in MOOCs has become essential (Hellas et al., 2018). Even though there are many reports on the prediction there is no prediction based on the features in ML using RF. Hence, this research was proposed.

In recent years, the augmentation of technology into education has helped in the success of students. Many systems have been developed to assist students in their learning and one such system is the Assessment and Learning in Knowledge Spaces (ALEKS). A combinatorial structure describing the possible states of knowledge of a human learner is

called a Knowledge Space. According to Knowledge Space Theory (Doignon & Falmagne, 1985) the breakdown of a mathematical domain to feasible subsets of mathematical concepts known to any individual with the knowledge of that domain. Further, Knowledge Space Theory is used in ALEKS to present the problems that the student is interested in learning. A screenshot of the ALEKS is shown in Figure 1.

Figure 1: *Screenshot of Aleks*

A large volume of data can be gathered and apprehended from MOOCs platforms during student interaction with learning activities including viewing of video lectures, the undertaking of quizzes, posting in discussion forums, and interacting with the courseware (Qiu et al., 2016; Kloft et al., 2014). Data captured from MOOCs can furnish valuable information for educators by probing the patterns present in the behavior of learners (Ramesh et al., 2014; Kloft et al., 2014).

The way the order of topics is shown and then tested follows a prerequisite structure, where some of the topics need to be mastered before progressing on to other topics in a systematic way. Though this is similar to any computer-based training (CBT)

system, ALEKS is more complex than other systems in using Bayesian networks to adaptively select the topics for the students.

Moreover, these networks of Knowledge Space Theory attempt to fill learning deficits and correct misconceptions adaptively and dynamically. It has been shown that ALEKS outperforms the highest quality teaching system designed and provided by educational experts (Craig et al., 2013). The experiment conducted using ALEKS and the results obtained proved the positive impact of ALEKS on students (Canfield, 2001). There is also a good correlation between the time spent on ALEKS and the learning outcome of the student (Stillson, & Alsup, 2003). Even though ALEKS exhibit positive results, it is a technologically intimating system and hence may not have the same results when used by a large group of students.

In this research time in the course is considered to be different from the time in real life. This way a student has the freedom to work for a day on the course and not come back for a week, but after a week the same student can return and work for another day on the course. In this case, the number of days the student worked on the course is not 2 weeks but 2 days. However, the student cannot enjoy this leverage beyond a period; for example, if there is an assessment period. Hence in this research, if a student has not returned to the course for more than three months, then that student is considered to have dropped out of the course.

In the MOOC course under consideration for this research, a student gains a topic in their knowledge space based on their performance in recurring tests. Therefore, the time unit in this research is the day of study under consideration of the course and aggregate it to a sequence as the student masters the topics shown as the learning progression of a

student. Further, this is the learning progression data, which we hypothesize can help to predict their day of dropout. Furthermore, given a set of features of changes in a student's learning progression from the start of the course till the given day of consideration, we can predict whether the student will continue further into the course or drop out of the course on this day of consideration. This research has significance because of the following reasons.

This research is focused on predicting the dropout of students from MOOC with the help of ML in particular by the application of RF using the 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 the day, a student drops out of the MOOC course helps in designing a relevant 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

CHAPTER 2

RELATED WORK

2.1. EDUCATIONAL DATA MINING AND LEARNING ANALYTICS

Educational Data Mining (EDM) is the application of data mining techniques to educational data to obtain solutions to problems in the field of education (Baker & Yacef, 2009). EDM engrosses the use of statistics, visualization, and machine learning techniques for the assessment and evaluation of educational data (West, 2012). Some of the EDM applications include the formulation of e-learning systems (Baker & Yacef, 2009; Lara et al., 2014), clustering educational data (Chakraborty et al., 2016), and making predictions of student performance (Chauhan et al., 2019). Several techniques are currently popular in educational data mining such as sequential pattern, clustering, prediction, classification, machine learning models, and association rule analysis (Salloum et al. 2020; West, 2012; Al-Shabandar et al.,2017)

EDM is a developing field focusing on applying statistical and ML techniques to analyze big educational data for a larger perception of student's behavior patterns and the learning environments. Several EDM studies have been carried out using ML techniques to determine various features including learners' performance, dropout, engagement, and interaction that meaningfully impact online learning platforms.

Dropout in MOOCs refers to the event of students failing to complete the course (Liyanagunawardena et al., 2014). Even though there are a lot of reports on the prediction of student dropout in MOOC, it remains the most important problem in this research area (Hellas et al., 2018). One of the reasons for this problem still being so important is that there has been no universal technique to predict student dropout that can be applied to

multiple courses. This is because the prediction models that exist are specific to a type of problem and are not diverse to be applied to various courses. Further, feature engineering is emerging as an important technique and the incorporation of features including test grades within the course could prove to be a useful and effective solution to the prediction problem in EDM (Dalipi et al., 2018).

Learning Analytics (LA) is an emerging field of research that intends to improve the quality of education (Baker & Siemens, 2014; Fiaidhi, 2014) LA is an analytics methodology oriented towards the evaluation, and extraction of comprehensive information about the learner from various features, such as cognitive, social, and psychological facets, to help the decision-maker reason about the learner's success and failure (Baker & Siemens, 2014; Fiaidhi, 2014). There are various techniques exploited by researchers in LA like Web analytics, Artificial Intelligence, and Social Network Analysis (Baker & Siemens, 2014). The key feature of LA is its capacity to evaluate actionable data in a more objective way (Fiaidhi, 2014, Gašević et al., 2014) The evaluation of such big data will help educators in deriving inferences about student performance with greater insight (Gašević et al., 2014). 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 (Qiu et al., 2016).

2.2. FEATURE ENGINEERING

There are many reports on the collection of features for the prediction of student's performance at the end of the course. Those results are useful in identifying the significant

features that reveal the student's performance but will not be useful to predict student dropout and failure.

Several studies aim at evaluating features that are created from learner's online activities (Jayaprakash et al., 2014; Márquez-Vera et al, 2016; Palmer, 2013), but few papers also use demographics features to perceive their influence on the study behavior of learners (Papamitsiou & Economides, 2014; Peña-Ayala, 2014). Early, the features considered for analysis include study time, study duration, content type, and features derived from social interactions. The emergence of the online learning platform as a stable and interactive platform transformed the features to assessment scores, assignment scores, clickstream, online forum interaction, and location for the analysis process (Zacharis, 2015). The selection and identification of significant features are some of the challenges for researchers due to diversity in platforms including MOOCs.

The role of demographics features was investigated by many to analyze the student's rate of retention in the course (Cen et al., 2016; Mueen et al., 2016; Huang & Fang, 2013; Marbouti et al., 2016). Around 120 types of a dataset of undergraduate students learning economics and business were analyzed and compared (Tempelaar et al., 2015). These different features including educational background, clickstream data, assessment scores, entry test scores, and learning personality data on students' performance were analyzed. Even though most of the studies focusses on finding the impact of key features on students' performance, there are also studies (Kizilcec et al., 2017; Kuzilek et al., 2015; Wolff et al., 2013) that concentrate on early prediction, intervention, learned support and appropriate feedback to guide and prevent student's dropout. There are several studies carried out at Open University, UK (Hlosta et al., 2018; Cui et al., 2020) to identify the

performance of students by analyzing several predictor features. These studies correlated the study behavior of learners to predict

1. Poor performance (when performance is less than a threshold value)
2. The success of students at the end of the course.

These investigations also showed that demographic features employed along with students' behavior feature offered enhanced predictive models in terms of performance and accuracy.

In an attempt to identify features which, compel students to drop out of the course, the features were classified into three categories (Lee & Choi, 2011).

1. Students' demographic features
2. Features related to course structure and requirement including the number of assessments, institutional support, interaction, difficulty level, and time duration.
3. Features related to environmental factors like the technology used, location, external noise, work environment, home environment, etc.

LA techniques on features generated from online courses and their impact on the prediction of students' performance were studied. The results showed that students performing well have a better engagement percentage compared to the students of poor performance (Soffer & Cohen, 2019).

The clickstream features including the online engagement of students are more accurate, objective, and comprehensive than self-reported data in measuring student's learning behavior (Winne, 2010). They are more authentic as it is collected from reliable learning environment while the learning behavior is generated from self-reported data.

Further, clickstream data are discreet and did not require student's full attention as they can be collected effortlessly without interfering with students' learning process (Sha et al., 2012). Furthermore, automatically collected clickstream data can offer large-scale and timely measures of students learning behavior which might help instructors in identifying the students' online engagements each day.

Recently, many studies have been performed to investigate clickstream data created from online learning platforms including MOOCs, Learning Management System (LMS), and Virtual Learning Environment (VLE) to evaluate students' online engagements. Although most of the studies try to investigate the relationship between clickstream data and students' online engagements, very few studies have gone one step further to enable instructors to know how and when to intercede students at the optimal time e.g., (Baker et al., 2019; Cicchinelli et al., 2018; Lim, 2016; Park et al., 2017).

In an attempt to use ML with different features, including several events, certified or not, days active, chapters explored, that are responsible for student retention in e-learning, the Decision Tree (DT) algorithm was used to establish the significant features that assist MOOC learners and designers in developing course content, course design, and delivery (Gupta & Sabitha, 2019). Various data mining techniques were applied to three MOOC datasets to evaluate the in-course behavior of the online students. Further, it is also claimed that the models used could be beneficial in the prediction of significant features to lower the attrition rate.

2.3. MACHINE LEARNING

Machine Learning (ML) is a proficient technique that can be applied to Learning Analytics with the ability to discover hidden patterns of student interaction with MOOCs. Machine learning has an advantage over conventional forms of statistical analysis, engaging importance on predictive performance over provable theoretical properties and priori super-population assumptions (Qiu et al., 2016). Moreover, an important feature of machine learning is its ability to analyze complex non-linear relationships from complex input features (Baker & Siemens, 2014; Gašević et al., 2014). Among the many machine learning techniques, the supervised RF machine learning technique has been employed to predict the student's dropout in the MOOCs platform.

The 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 (Costa et al., 2017). Earlier studies have indicated that the development of predictive models, but many challenges limit their application to a specific learning platform. Creating of predictive and flexible models which can adapt and/or adjust in the different learning environment is a big challenge. The limitations were the presence of various course structures, different instructional designs, and diverse online platforms (Gašević et al., 2016).

In recent years, researchers have used both statistical and predictive models to explore in a large repository including formal and informal educational settings (Bozkurt et al., 2018; Baker & Inventado, 2014).

The earliest prediction of the dropout students taking online courses using a time-series clustering approach was reported. In which, the time-series clustering approach developed predictive models was of better accuracy than the conventional aggregation method (Hung et al., 2015).

An early warning system that employed students' eBook reading data to predict students' performance for academic failure was developed (Akçapçnar et al., 2019). In this 13 ML algorithms were employed to train the model using data from different weeks of the semester. Among them, the best predictive model was selected based on the accuracy/Kappa metric (Cohen, 1960) and recommending optimal time for the instructors to intervene (Adnan et al., 2021). Moreover, all predictive models improved their performance results when weekly data was used progressively during the training method. The early warning system predictive models were successful in classifying low and high-performance students with an accuracy of 79% starting from the 3rd week. The RF outperformed other algorithms when the complete 15 weeks data were investigated by different algorithms, but the J48 outperformed all other algorithms when evaluated with the transformed data. Finally, the Naïve Bayes (NB) showed better performance when using categorical data.

Predicting the performance of the students early in the course is an exciting problem in online learning environments due to diversity in course structure and MOOCs design. While the popularity of LMS/MOOCs is rising, there is a necessity for an automated intercession system that might deliver timely feedback to students. To incorporate an automated intervention system with LMS/MOOCs, researchers have implemented different ML algorithms which, can support instructors in delivering educated support to

students during the learning process. ML algorithms including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees (DT), and RF are generally trained using daily, weekly, or monthly students log data to discover students' learning patterns. Deep learning (DL) algorithms are also employed in generating predictive models because they can handle raw data directly. The Recurrent Neural Network (RNN) algorithm trained on raw log student records was used for the prediction of students learning performance at the end of the course (Kőrösi & Farkas, 2020). In this, RNN in providing superior performance as compared to standard baseline methods.

Cano & Leonard (2019) 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 the at-risk student to improve their study performance. The genetic programming technique works nicely with multi-view learning. The prediction model learned and evolved is self-explaining without further adjustment. Moreover, while operating genetic programming approach results in the natural evolution of the classification rules evolved with the availability of new data. The early warning system built with comprehensible Genetic Programming classification rules specifically aims at underperforming and underrepresented students. Understandable feedback is offered to students, instructors, and administration staff utilizing three interfaces to offer timely support to students to hold them on the right track. The main shortcoming of this analysis was that the author did not mention the different semester stages at which the performance metrics such as accuracy, sensitivity, specificity, and Kappa were computed using a multi-view genetic programming algorithm along with other machine learning algorithms.

The ML algorithm, logistic regression a was employed to identify students who are liable to drop out in an e-learning course (Burgos et al, 2018). In this, the history of student's grades was employed as an input to model the performance of students. Further. this technique showed a higher performance score invalidation 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. Furthermore, the tutoring action plan based on logistic regression reduced the dropout rate by 14%.

Knowledge discovery in databases (KDD) was employed to mine information that may enable teachers in finding the interaction of students with e-learning systems (Lara et al., 2014). This method creates reference models which can be employed to predict the student's dropout in classes. The other technique, System for Educational Data Mining (SEDM), evaluates two groups of students for a single course i.e., dropout students who are not permitted to sit in the final examination and non-dropout students who are qualified to sit in the final examination. SEDM was able to generate study patterns for both groups, which may be beneficial for instructors to explain students' study performance.

These studies help in the prediction of student's performance including dropout of the course, however, none of these studies predicts students at-risk of dropout at a different percentage of course length. Further, there is no study on the prediction of the dropout of students using RF with features explained in this research. Hence this research was proposed.

CHAPTER 3

DATA DESCRIPTION

This section describes the students in the data considered for this research. This data cannot be made available publicly 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 gives us an idea of the background of the students and such a description helps us in understanding the impact of this research. Table 1 shows the distribution of students in this course.

***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 can 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. This disparity between students who complete the course exists across all demographics. Table 2 shows this disparity exists in different age categories. This dataset has students from the age of 9 to the age of 70 as shown in the table.

Table 2: *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

The disparity between students who complete the course and students who drop out can also be seen between the gender of students in the dataset. This is shown in the Table 3.

Table 3: *Distribution of Students Across Different Gender Groups*

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

The same amount of disparity is also seen among students in different ethnic groups in this course. This can be seen in Table 4 that shows this higher number of students who drop out when compared to students who succeed in each ethnic group found in this course. As mentioned before this problem exists in all demographics of students and this shows that this student dropout is a major problem that needs to be addressed. Since this problem is seen to exist in all groups of students. Hence, all students can be considered in this problem without any grouping of students. This is also one of the main motivations of this research since the solution to this problem helps all students irrespective of their background.

Table 4: *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

The methodology performed in this research to solve this problem is explained in the following sections

CHAPTER 4

METHODOLOGY

This chapter will explain the methodology of Machine Learning modeling to predict the dropout of students in MOOC. It is organized into three parts data handling, machine learning modeling, and model evaluation. The flow of the methods is explained in Figure 2.

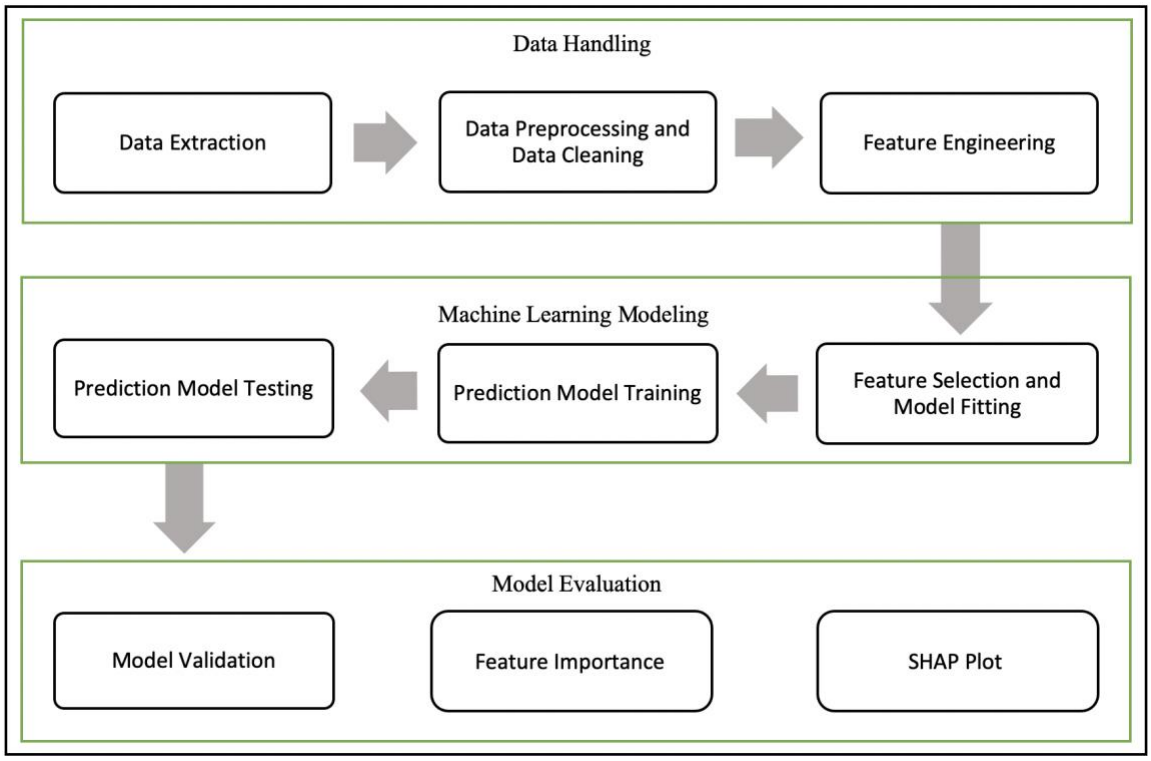

Figure 2: *The Flow of Methodology*

4.1. DATA HANDLING

Data handling is the technique to get the data from the data source to the machine learning model. This is done by the standard process of KDD (Fayyad *et al.*, 1996). This part is segmented into three sections as depicted in Figure 3 and at the end of the third section, the data will be ready for modeling.

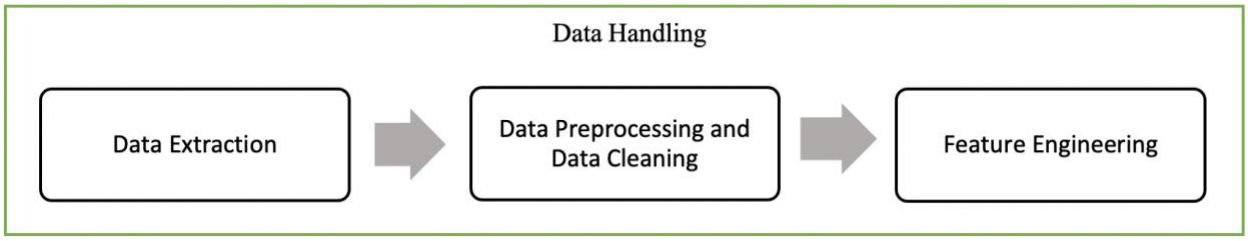

Figure 3: *The Flow of Data Handling*

4.1.1. DATA EXTRACTION

Data Extraction involves extracting data from the data source by using data mining methods. 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) which uses the ALEKS system. It consists of students taking this course starting from March 2016 to March 2020. The ALEKS platform has an API, which is used, in this research, to get the data into the SQL Database. I used the data selection method here which is to use a select query on the data source. This method works for most of the databases in general and well for SQL databases in particular (Fayyad et al., 1996). SQL is the type of database used in this research. Once the data selection is performed, the queried table is stored as a comma-separated values (***CSV***) file in either a local machine or on a cloud that is accessible for the experiment. This experiment is performed on a Jupyter notebook using an easily accessible CSV file and processed by python language.

The course under consideration is self-paced and hence there is no specific time or schedule to complete this course. The total number of participants in this course is 3172 and they will have some 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. Among the 3172 students, only 396 students have completed this course while the remaining 2776 students have not due to some reason. This shows that only 12.5% have completed the course successfully and 87.5% of students in this course have dropped out.

The data is in 4 different reports:

- class_report,
- assessment_report
- progress_report
- timeandtopic_report.

The class_report is the highest-level data that has not been used for this research while the assessment_report, progress_report, and timeandtopic_report have been grouped with student ID as the key. The attributes considered from the three datasets are tabulated in Table 5.

Table 5: *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

Once we have these target CSV files from the data source, we move onto the next step of data preprocessing.

4.1.2. DATA PREPROCESSING AND DATA CLEANING

Data Preprocessing is the process of extracting data that is needed for the machine learning model from the target data. The features of changes in student learning that are associated with the student dropout were found and used to predict this event. The extracted dataset has data as a time series with no primary keys. Since we need data with each student as the key and the data has to be mapped to the machine learning model, we need to perform grouping based on student identity. Research on a similar EDM dataset performed their experiment by grouping their data based on student identity as a preprocessing step which helped give structure to the raw dataset (Saa, 2016). A similar grouping is performed in this research. Once grouped we can extract the learning progression data for each student from the target data and store it as a table called the preprocessed data.

The student in ALEKS goes through a knowledge check after every topic or every 120 minutes 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 don't clear, then they lose these topics. So, the topics mastered in this research are the event of either mastering the topics or not in these knowledge checks on a day it is occurring within the course. 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 by the instructors. This research focuses on predicting the dropout occurrence, so we consider the topics mastered of these 2776 students. The data table holds a student and their entire learning information in one row. Here, we take 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 have 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 have this attribute measured for each day. This attribute is a cumulative score, and Figure 4 shows the rate of learning progression of each student.

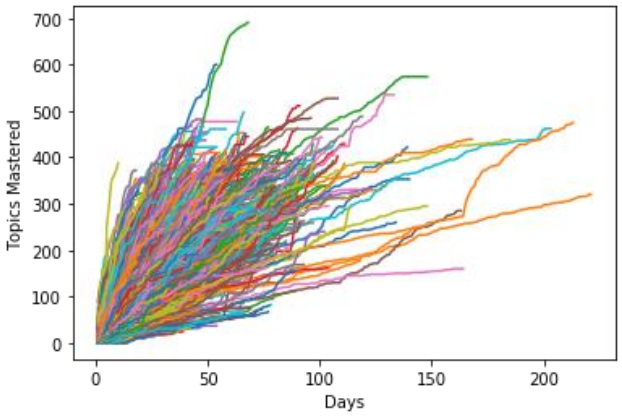

Figure 4: *The Cumulative Rate of Learning Progression of Each Student*

For this research, we need to find the changes in the learning progression. To find these changes in learning progression, we calculate the topics mastered by the student for each day. It is done by calculating the difference between the progress of that day with the total progress made by the student before that day for each day and this gives the topics mastered by a student on each day. These changes in the learning progression of a student are the dataset from which we obtain multiple features to be used by our machine learning model. Figure 5 visualizes the rate of changes in learning of each student.

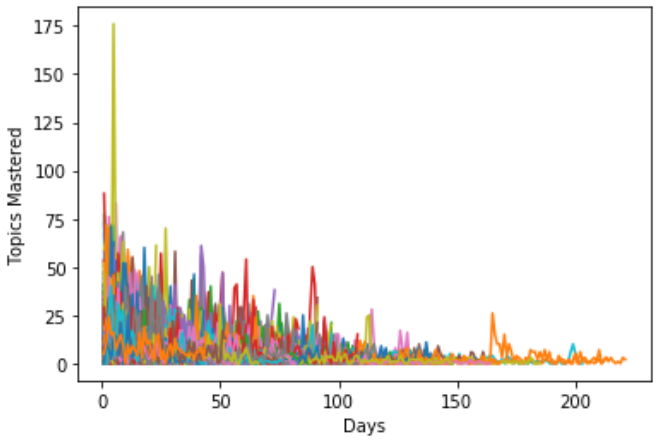

Figure 5: *The Rate of Changes in Learning Progression of Each Student*

Sequence classification is a predictive modeling problem where you have some sequence of inputs over space or time and the task is to predict a category for the sequence. (Fei, & Yeung, 2015). There are a lot of Self-paced MOOC courses and they do not offer a cohort structure or completion time requirement and thus this time unit aggregation becomes complex. This preprocessing approach as depicted in Figure 6 helps in simplifying this complex data aggregation. The data needed for this experiment is derived from the learning progression data, which is the *topics_mastered* in *timeandtopic_report*.

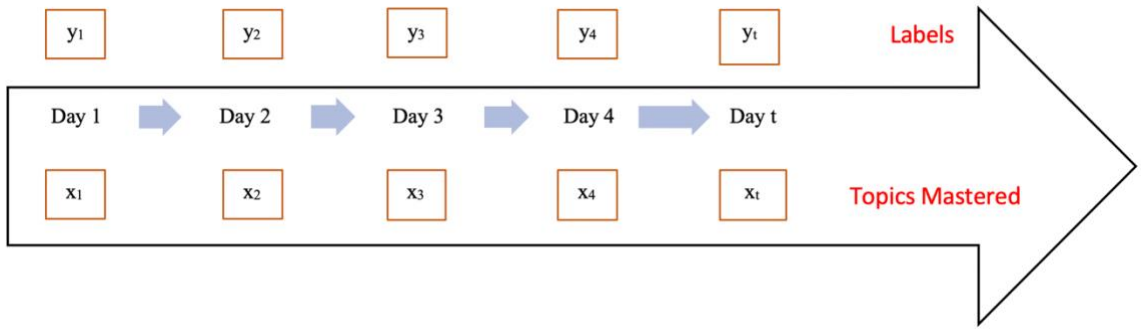

Figure 6: *The Data Preprocessing Method*

The preprocessing method proposed by Fei, & Yeung (2015) is used in this research. 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. Once we obtain the preprocessed data, we need to clean it for the modeling. Although the preprocessed data has all the necessary information, it cannot be used for modeling as it is structured but raw. Most of the attributes of the preprocessed data are of different data types and so we clean this data to be of the same numerical datatype for modeling. The data cleaning will result in the transformation of preprocessed data into cleaned data ready for the machine learning model (Fayyad *et al.*, 1996).

4.1.3. FEATURE ENGINEERING

The cleaned data in the experiment is now used to generate the features that can be used in the model. The features generated in this research describe the rate of student learning from the start of their course to the days of consideration on the course. The dataset is transformed into a feature table and that each row depicts the rate of student learning on any given day from the start of their course. Further, the 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 continues in the course.

From the cleaned data, the topics mastered on each day the student learns during their course are represented by the array of *topics_mastered* in this research. 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 (Nanopoulos et al., 2001). An example of rate of changed in student learning is shown in Figure 7.

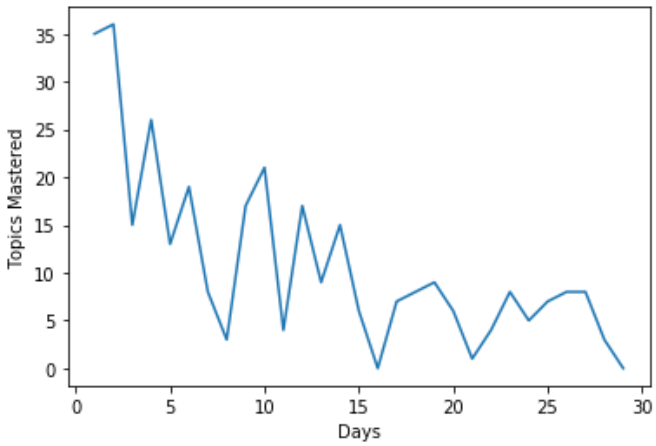

Figure 7: *An Example of Rate of Changes in Student Learning*

Feature engineering is the process of generating values to represent raw unstructured data. Here this time series varies in length and shape for every student and feature engineering is performed to represent this data on each day for each student.

Average is used to represent the central value of the entire given rate of learning. In Figure 7, 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 7 is very rough four features namely, skew, standard deviation, variance, and kurtosis, are used to represent this roughness.

The standard deviation is a measure of deviation from the mean of a given time series. It is calculated by using the formula,

$$Standard\ Deviation = \sqrt{\frac{\sum(x_i - Mean)^2}{Size\ of\ the\ given\ array}} \quad (1)$$

Where x_i is the i^{th} topics mastered in the given array.

The variance is a measure of inconsistency in the time series or how to spread out the values are from the mean. It is calculated by squaring the standard deviation.

Skewness, kurtosis, and length of distribution are features that explain the shape of the distribution. Skewness is a measure of asymmetry and is used to represent the lack of symmetry in a given distribution. Pearson's second skewness coefficient (Doanne & Seward, 2011) is used in our experiments. It is calculated by the formula,

$$Skew = \frac{3(Mean - Median)}{Standard\ Deviation} \quad (2)$$

Kurtosis is a measure of the sharpness of the peaks in the distribution. It is calculated by dividing the fourth central moment by the standard deviation (Doanne & Seward 2011). We also consider the length of the distribution to better represent the shape of the distribution.

There is one more characteristic of these topics_*mastered* distributions that are taken into consideration. The relationships of *topics_mastered* values given in the distributions are used in this research. The slope between two points of distribution provides a measure of the relationship between these two points (Akima, 1970). In this research, we want to see the relationship between the first day in the distribution and the last day given in the distribution, and if that relationship is different from the relationship between the last two days in the distribution. The reason for seeking these relationships is to map the last activity of the student to their overall activity and changes seen in this relationship could be predictive of student dropout. The relationship between the first and the last day in distribution is calculated as overall trajectory using the formula,

$$Overall\ Trajectory = \tan^{-1}\left(\frac{Topics\_mastered_n - Topics\_mastered_0}{n}\right) \quad (3)$$

Where n represents the length of the given array of topics_mastered or the day of consideration, $Topics\_mastered_n$ represents the topics_mastered on the day of consideration and the $Topics\_mastered_0$ represents the topics_mastered on the first day of the course for the student. The relationship between the last two days in distribution is calculated as the final trajectory using the formula,

$$Final\ Trajectory = \tan^{-1}\left(\frac{Topics\_mastered_n - Topics\_mastered_{n-1}}{1}\right) \quad (4)$$

Here the denominator is 1 because the days considered are a day apart, the $Topics\_mastered_n$ represents the topics_mastered on the day of consideration and the $Topics\_mastered_{n-1}$ represents the topics_mastered on the day before the day of consideration. Thus, in total this research considers eleven features to represent the rate of student learning in the MOOC course till the day of consideration for analysis as shown in Table 6.

Table 6: *Features Engineered in this Research*

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

These 11 features make the learning features of a student on a day. A small sample of the final feature table is shown in Table 7.

Table 7: *A Sample of Feature Table*

Moving average 2	**Moving average 3**	**Moving average 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

After these features are engineered, we can move onto the feature selection process in the ML modeling phase.

4.2. 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 as shown in Figure 8.

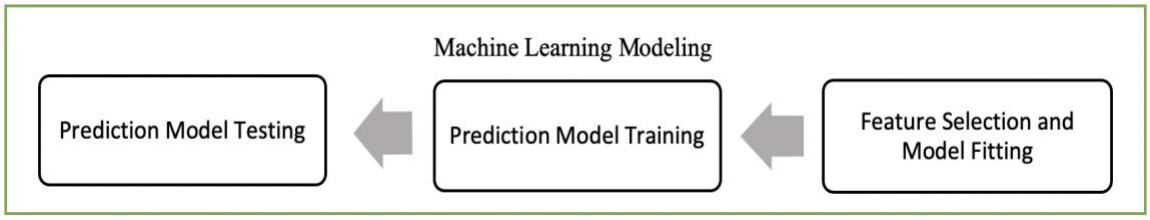

Figure 8: *The Flow of Machine Learning Modeling*

4.2.1. FEATURE SELECTION AND MODEL FITTING

In this step, the features generated in the previous step are evaluated and validated. To predict the student learning outcomes in MOOC, Exploratory Data Analysis (EDA) technique called the correlation matrix method is used to validate the features (Al-Shabandar et al., 2017). This method helps in removing features with dependencies with each other. This in turn would avoid data leakage in the ML model and increase the success of the modeling experiment. This is an iterative process and is to be repeated until the features are independent of each other. The features are fairly independent when a correlation value is less than 0.5 and strongly dependent if the correlation value is 0.8 or greater (Hall, 1999). Thus, the features with no dependency are derived from the correlation matrix and used for modeling. These features are now converted into the input

vector and the target column is converted into the output vector. However, the data is not balanced at this stage. The target spread is shown in Table 8.

Table 8: *Target Values*

Target value	Number of data points
0	39529
1	2776

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 of continue while 6.5% of the data points are of dropout. Different ML models predicting student dropout in MOOCs also faced this type of imbalance and the usage of Synthetic Minority Oversampling TEchnique (SMOTE) helped in overcoming this imbalance to complete the modeling (Hong et al., 2017; Wang et al.,2006). SMOTE creates synthetic data for oversampling the minority class. SMOTE uses a version of the k-nearest neighbor algorithm to create the synthetic data. The nearest neighbors are calculated, and then synthetic data is generated between randomly selected nearest neighbors (Chawla et al., 2002). This not only randomizes the data to eliminate bias but also creates data while maintaining the same data space. This is one of the main reasons SMOTE is used as a solution for big data imbalance problems and is well accepted in the research community (Hong et al., 2017; Wang et al.,2006). In this research, we have also employed SMOTE to overcome this imbalance. The balance in the data after the application of SMOTE is shown in Table 9.

Table 9: *Target values after SMOTE*

Target value	Number of data points
0	39529
1	39529

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

4.2.2. PREDICTION MODEL TRAINING

The most commonly used ML models in EDM include XGBoost, RF, and SVM (Adnan et al., 2021). Among the ML models in EDM, RF performs better (Adnan et al., 2021, Al-Shabandar et al., 2017). The RF is an ensemble ML model that creates many decision trees during model training and ranks them based on the maximum number of correct predictions. Then the RF corrects these decision trees from overfitting the training data. Once the RF model is trained, the testing data is fed as input and the RF gives the prediction. The relevance of input is seen by the mode of classes with the highest votes. The accuracy of the RF model is as good as or sometimes even better than most of the ML models (Breiman, 2001; Alamri et al., 2019). 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. The Gini impurity is a measure of the number of randomly given feature sets that would be incorrectly labeled if the tree is randomly

labeling the data based on the label distribution in the dataset. 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 a lot of outliers. The dataset in this research, although goes through proper feature engineering and feature selection processes, still holds a lot of outliers and hence random forest is better suited for this research. The 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 (Archer & Kimes, 2008). This research focuses on the feature engineering approach and its contribution and hence making RF the prime choice for the machine learning model. Here, the RF model is trained with the training data and is 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 are fed.

4.2.3. PREDICTION MODEL TESTING

Once the model is trained, we perform experiments to test the model for its performance and see if the model can predict the correct outcome for a given set of features. In this study, we have performed 5 experiments to test the performance of the model with different sets of inputs. The experiments are 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 5 experiments are explained in the Experimental Setup section. Once we get the results from these experiments the model is evaluated as described in the model evaluation section.

4.3. MODEL EVALUATION

There are three main processes of model evaluation employed in this research. They are model validation, feature importance, and SHAP plot. The six experiments we performed will result in different model predictions and model validation is performed to evaluate the results of these six experiments. The feature importance and the SHAP Plot are visualizations used to identify the most important contributor to the model's predictions.

4.3.1. MODEL VALIDATION

Different methods have been developed to validate the models including direct correlation, Cohens Kappa, and Accuracy (Bradley, 1997) but accuracy is not recommended for evaluating the model because it depends on the base rates of different classes (Algarni, 2016). 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 that considers both true and false classification results based on precision and recall is the better metric. There is always a positive class and a negative class in a classification model. The positive class represents the prediction outcome, and it can be a dropout or continue based on the experiment performed. There are four values used to calculate the metrics to validate the model. *True Positive* is the number of correct positive class predictions, *True Negative* is the number of correct negative class predictions, *False Positive* is the number of incorrect positive class predictions and *False Negative* is the number of the incorrect negative class prediction We have also performed the model validation using these four metrics (Goutte & Gaussier, 2005).

Precision is defined as the number of successful positive predictions that belong in the positive class. Precision is calculated by the formula,

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (5)$$

The recall is defined as the number of actual positive class data points that were predicted to be in a positive class. The recall is calculated by the formula,

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (6)$$

The F1-score is the combined method of Precision and Recall considering both true and false classification results. F1-score is calculated by the formula,

$$F1\ score = 2\ v\acute{o}\ \frac{Precision\ v\acute{o}\ Recall}{Precision + Recall} \quad (7)$$

Even though accuracy alone cannot be used to validate a model, it still portrays the performance of the model and hence 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. This curve is obtained by plotting the True Positive Rate to the False Positive Rate, where True Positive Rate is the Recall calculated before and False Positive Rate is calculated by the formula,

$$False\ Positive\ Rate = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (8)$$

With this curve, we can also see the Area Under the Curve. The Area Under the Curve (AUC) is the other validation method of evaluating a prediction model. In this research, we have also studied the AUC to validate the method. A value of at least 0.7 for these metrics is accepted in the research community.

These are the set of metrics used in this model validation method to evaluate the machine learning model.

4.3.2. 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 (Sharkey, & Sanders, 2014). 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.

4.3.3. SHAPELY ADDITIVE EXPLANATIONS (SHAP) PLOT

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. Further, these plots were used to evaluate the model and the features to predict engagement in video lectures (Bulathwela et al., 2020). 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.

This is done by calculating the shapely values for the features. The shapely value is a gamification concept from the field of economics and the formula to calculate this value is shown below (Ichiishi, 2014).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] . \tag{9}$$

The above formula is used by the python library to decompose the prediction of the model to find the impact that each feature has on that prediction. This is visualized by the SHAP plot. It is obtained by comparing the model's prediction including and excluding the feature. In this research, the shap python library is used to visualize the shapely values for the features used. In the SHAP plot, the impact of the features on the prediction of the model is observed.

These evaluation methods have been used extensively in the industry and by the research community and hence it has been used for the prediction of a student in a MOOC course on their continuation or dropout in the course under consideration from the changes in the learning features on any day. In summary, it is proposed that the above-discussed methodology can be used to predict whether a student will drop out or continue the MOOC course based on the changes in learning behavior of the students. The findings of this methodology are discussed in the following section.

CHAPTER 5

EXPERIMENTAL SETUP

5.1. FEATURE SELECTION

The feature selection process on the data under consideration is done as per the methodology discussed in the feature engineering section 4.1.3. The correlation between the 11 features is established and is shown in Figure 9.

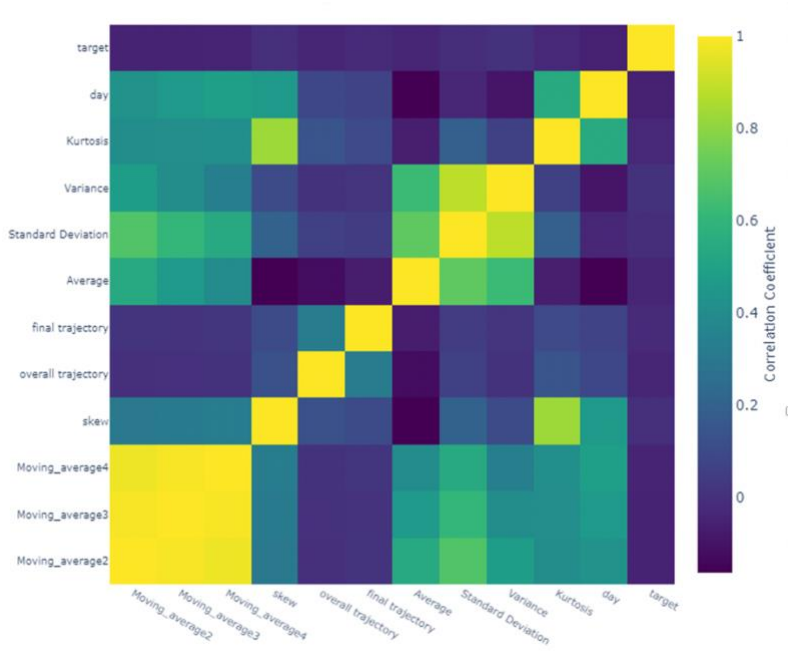

Figure 9: *Correlation Matrix of Features*

Figure 9 shows that 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 3. kurtosis and skew. The dependency value between kurtosis and days feature is seen to fall 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 10.

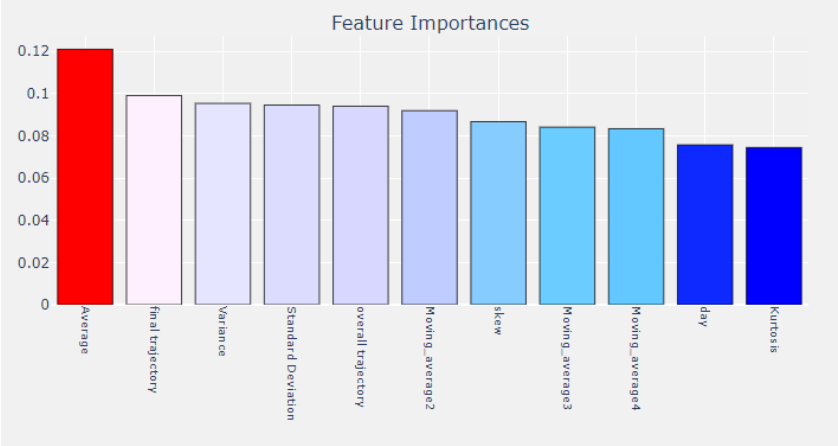

Figure 10: *Feature Importance Plot*

From Figure 10, the most important feature in each of the three dependent feature groups is 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 is tried and the correlation matrix obtained after feature selection is shown in Figure 11.

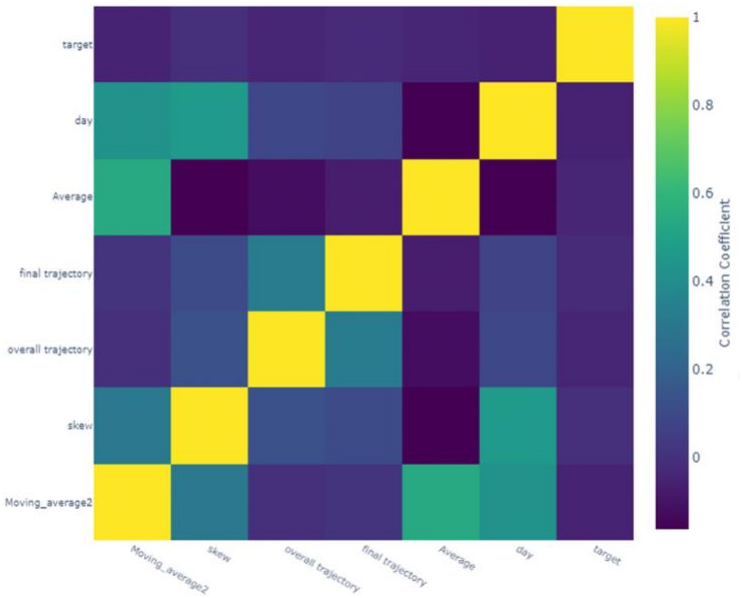

Figure 11: *Correlation Matrix of Features after Feature Selection*

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. 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 is used for training the model and the remaining 25% of the data is used to test the model.

5.2. MODEL TRAINING

The RF model was trained from scikit-learn (https://scikit-learn.org/) with the specifications shown in Table 10.

Table 10: *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 is used to obtain the feature importance plot as shown in Figure 15 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 is trained with these specifications shown in Table 5, model testing is performed.

5.3. 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 11.

Table 11: *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.

CHAPTER 6

RESULTS AND DISCUSSIONS

6.1. EXPERIMENTAL RESULTS

This experiment involves testing the model with the entire testing data separated from the training data before training the model. We pass the input testing data to the model and receive its predictions for this set of input. Then we check the output testing data values with the model's prediction values and use them to calculate the model validation metrics. The results of the model validation are shown in Table 12

Table 12: *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 the 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 12 shows the result of the ROC curve analysis performed for the model trained and tested in this research and it can be seen that the AUC is far away from the 0.5 line and this means that the model has covered the dataset well and can predict the student dropout or continue for most cases in the dataset.

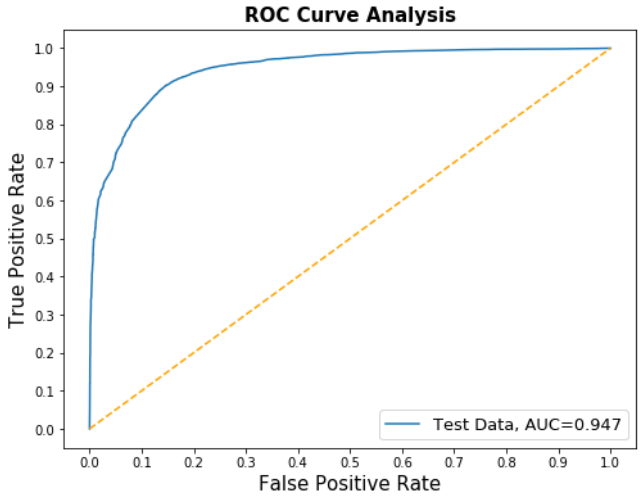

Figure 12: *The ROC of the Model*

This research investigated the performance of the model further. The testing data was segmented for each day of consideration and the model was tested and validated for each of these sets of data. This is done to visualize the performance of the model on different sets of data. The accuracies of the model on different days are seen in Figure 13. It can be observed that the accuracy of the model is consistently above 70% and mostly above 80%.

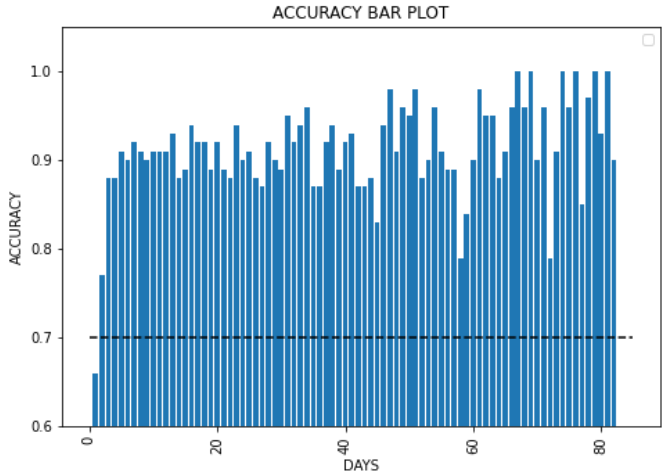

Figure 13: *Accuracy of the Model on Different Days*

The precision of the model on different days can be seen in Figure 14. It can be observed that the precision of the model is always above 70% and consistently above 80% and mostly above 90%.

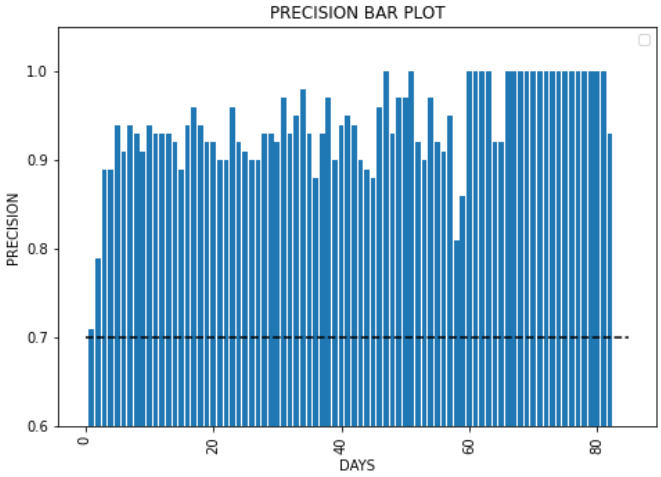

Figure 14: *Precision of the Model on Different Days*

The recall of the model on different days can be seen in Figure 15. It can be observed that the recall of the model is always above 80% and consistently above 90%.

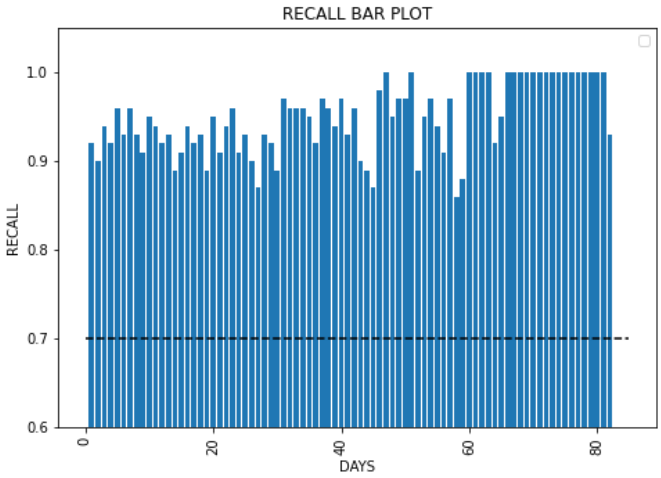

Figure 15: *Recall of the Model on Different Days*

The f1-score of the model on different days can be seen in Figure 16. It can be observed that the f1-score of the model is always above 70% and consistently above 80% and mostly above 90%.

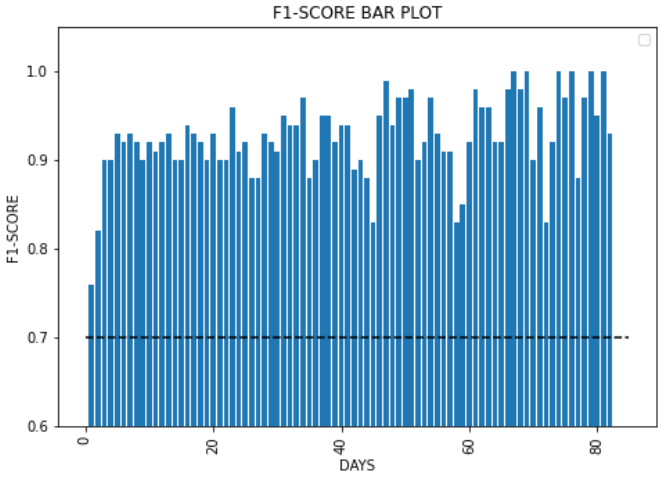

Figure 16: *F1-Score of the Model on Different Days*

These results show that the model is performing 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. But 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.

6.2. 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 17.

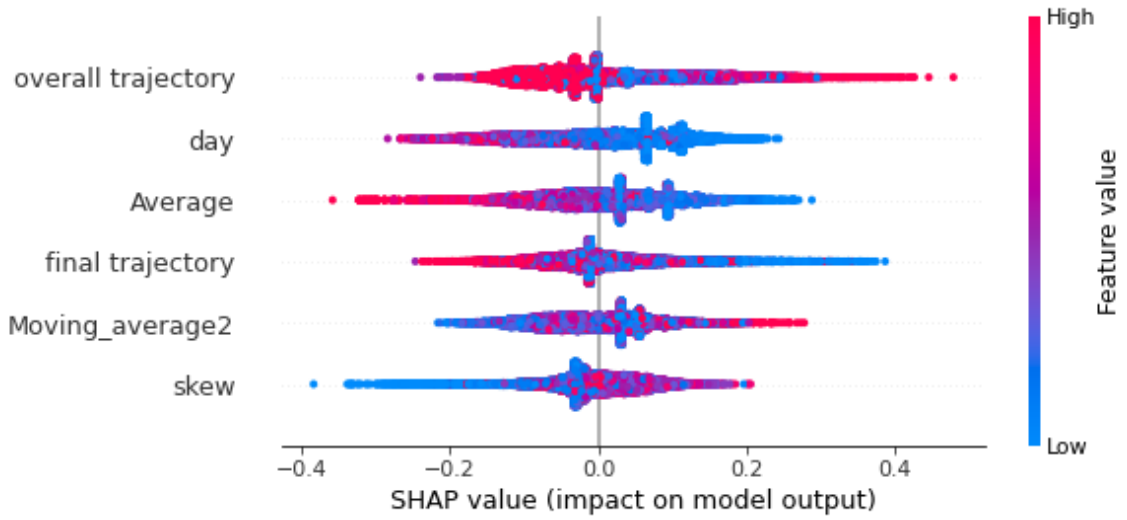

Figure 17: *The SHAP Summary Plot for this Prediction Model*

From Figure 17, 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.

To better understand the feature interactions, let us look at 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 13.

Table 13: *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 predicts dropout, class '1'. The shap force plot shown in Figure 18 helps visualize the feature interactions that lead to this prediction.

Figure 18: *SHAP Force plot 1*

In Figure 18, we can see that 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 14.

Table 14: *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 19 helps visualizes the feature interactions that lead to this prediction.

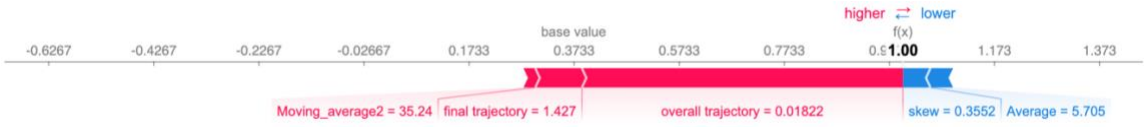

Figure 19: *SHAP Force Plot 2*

In Figure 19, we can 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.

6.3. LIMITATIONS

The features engineered from this data are common statistical and empirical measures of a sequence of data points. More complex and sophisticated features could prove to be better predictors than the ones used in this research.

This research is on one Math course and specifically on the course that uses ALEKS. With data from a similar group of courses, this research would have been able to predict the dropout of students across different courses.

Though this research uses one of the most fundamental attributes of data that could exist in many other courses, it is still specific to the ALEKS system. This limitation might cause this research to not be able to extend to other MOOCs as much as this research would hope otherwise.

The fundamental attribute of data mentioned above is the topics mastered by the student and this is a scoring given after a test. This makes this research dependent on the scorer. The effectiveness of the grader is directly related to the validity of this research. In the math course, the system evaluates the multiple-choice test, and as long as the system functions properly the grading is effective, but the same cannot be said for other courses. So this research is limited by the proper functioning of the course software.

This research predicts the student dropout on the given day because it is limited by the data and the complexity of the ML model. Early prediction of student dropout might be possible when these limitations are overcome.

6.4. SUMMARY

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 the 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% (Adnan et al., 2021), 86.5% (Hong et al., 2017), 87.6% (Alamri et al., 2019). This shows that if the learning progression data is 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 the 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 on the course.

CHAPTER 7

CONCLUSION

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 data set is analyzed using RF, the feature and modeling evaluation is done by Precision, Recall, F1 - score, AUC, 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%.

7.1. FUTURE WORK

There are multiple avenues to extend this research to solve the dropout problem by prediction. A few ideas that could work are discussed below.

This approach used for multiple ALEKS courses could prove that this methodology can predict student dropout in most of the ALEKS courses with topics mastered data. This would also open the avenue to solving the dropout problem across multiple types of courses on MOOCs which would increase the extensibility of research done in this domain.

A complex machine learning algorithm can be augmented into the approach, which could perform better than random forest to increase the accuracy of prediction and the performance of the model. It could also help when data across multiple courses are considered to perform prediction with high accuracy despite multiple different datasets and feature interactions.

Develop more complex sequence-based features which could help improve the model's performance and make the model ready to handle different course datasets. There are much time sequence-based features that are more complex than the statistical features used in this research and further research on more complex feature engineering may help in the earlier prediction of student dropout more accurately.

Develop an intervention after the student has dropped out and augment it to this prediction model and survey the impact it has on the student's learning. Further research on how different interventions on the day this research predicts the student will drop out, can help in establishing the impact of this research.

REFERENCES

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., ... & Khan, S. U. (2021). Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access*, *9*, 7519-7539.

Akçapçnar, G., Hasnine, M.N., Majumdar, R., Flanagan, B., & Ogata, H. (2019). Developing an early-warning system for spotting at-risk students by using eBook interaction logs. *Smart Learning Environments*, 6(1), 4.

Akima, H. (1970). A new method of interpolation and smooth curve fitting based on local procedures. *Journal of the ACM*, *17*(4), 589-602.

Alamri, A., Alshehri, M., Cristea, A., Pereira, F.D., Oliveira, E., Shi, L., & Stewart, C. (2019). Predicting MOOCs Dropout Using Only Two Easily Obtainable Features from the First Week's Activities. In: A. Coy, Y. Hayashi, M. Chang (Eds) *Intelligent Tutoring Systems*, (pp 163-173) Springer, Cham. https://doi.org/10.1007/978-3-030-22244-4_20

Algarni, A. (2016). Data mining in education. *International Journal of Advanced Computer Science and Applications*, *7*(6), 456-461.

Al-Shabandar, R., Hussain, A. Laws, A., Keight, R.., Lunn, J., & Radi, N. (2017, May 14-19). Machine learning approaches to predict learning outcomes in Massive open online courses. *International Joint Conference on Neural Networks*, (pp. 713-720). Anchorage, AK, USA. 10.1109/IJCNN.2017.7965922.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational statistics & data analysis*, *52*(4), 2249-2260.

Baker, R. S., & Inventado, P. S. (2014). *Educational data mining and learning analytics in Learning Analytics*, New York, NY, USA:Springer, pp. 61-75.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3-17.

Baker, R., Evans, B., Li, Q., & Cung, B. (2019). Does inducing students to schedule lecture watching in online classes improve their academic performance? An experimental analysis of a time management intervention, *Research in Higher Education*, 60(4), 521-552.

Baker, R.S.J. D., & Siemens, G. (2014). Educational Data Mining and Learning Analytics, In R. Keith Sawyer (Ed.). *Cambridge Handbook of the Learning Sciences*, 2nd Edition, (pp.253 – 274). New York, NY, Cambridge University Press.

Bozkurt, A., Yazıcı, M., & Aydin, I. E. (2018). Cultural diversity and its implications in online networked learning spaces In Information Resources Management Association

(Eds.). *Research Anthology on Developing Effective Online Learning Courses* Hershey, PA, USA:IGI Global, pp. 56-81.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145-1159.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Bulathwela, S., Pérez-Ortiz, M., Lipani, A., Yilmaz, E., & Shawe-Taylor, J. (2020, July 10-13). Predicting Engagement in Video Lectures. In Proceedings *International Conference on Educational Data Mining*,Ifrain, Morocco. arXiv:2006.00592**.**

Burgos, C., Campanario, M. L., Peña, D. D. L., Lara, J. A., Lizcano, D., & Martínez, M. A.(2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, 66, 541-556.

Canfield, W. (2001). ALEKS: A Web-based intelligent tutoring system. *Mathematics and Computer Education*, *35*(2), 152.

Cano, A., & Leonard, J. D. (2019). Interpretable multiview early warning system adapted to underrepresented student populations, *IEEE Transactions of Learning Technologies*, *12*(2), 198-211.

Cen, L., Ruta, D., Powell, L., Hirsch, B., & Ng, J. (2016). Quantitative approach to collaborative learning: Performance prediction individual assessment and group composition. *International Journal of Computer-Supported Collaborative Learning*, *11*(2), 187-225.

Chakraborty, B., Chakma, K., Mukherjee, A. (2016, March 17-18) A density-based clustering algorithm and experiments on student dataset with noises using Rough set theory. *IEEE International Conference on Engineering and Technology*, (pp. 431–436), Coimbatore, India. 10.1109/ICETECH.2016.7569290

Chauhan, N., Shah, K., Karn, D., Dalal, J. (2019, April 8-9). Prediction of student's performance using machine learning. *2nd International Conference on Advances in Science & Technology*, Mumbai, India.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Cicchinelli, A., Veas, E., Pardo, A., Pammer-Schindler, V., Fessl, A., & Barreiros, C. Lindstadt,S. (2018, March 7-9) Finding traces of self-regulated learning in activity streams. *Proceedings in 8th International Conference Learning Analytics & Knowledge*, Sydney, NSW, Australia, pp. 191-200.

Costa, E. B., Fonseca, B., Santana, M. A. de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of student's

academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247-256.

Craig, S. D., Hu, X., Graesser, A. C., Bargagliotti, A. E., Sterbinsky, A., Cheney, K. R., & Okwumabua, T. (2013). The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors. *Computers & Education*, *68*, 495-504.

Cui, Y., Chen, F., & Shiri, A. (2020) Scale up predictive models for early detection of at-risk students: A feasibility study. *Information and Learning Sciences*, *121*(3), 97-116.

Dalipi, F., Imran, A. S., & Kastrati, Z. (2018, April 17-20). MOOC dropout prediction using machine learning techniques: Review and research challenges. In *IEEE Global Engineering Education Conference,* Santa Cruz de Tenerife, Spain (pp. 1007-1014). 10.1109/EDUCON.2018.8363340

Doanne, D., & Seward, L. E. (2011). Measuring skewness. *Journal of Statistics*, *19*(2), 1-18.

Doignon, J. P., & Falmagne, J. C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, *23*(2), 175-196.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34.

Fei, M., & Yeung, D. Y. (2015, November 14-17). Temporal models for predicting student dropout in massive open online courses. In *2015 IEEE International Conference on Data Mining Workshop.* Atlantic City, NJ, USA (pp. 256-263). 10.1109/ICDMW.2015.174

Feng, W., Tang, J., & Liu, T. X. (2019). Understanding Dropouts in MOOCs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 517-524. https://doi.org/10.1609/aaai.v33i01.3301517

Fiaidhi, J. (2014). The Next Step for Learning Analytics. *IT Professional*, 16 (5). 4–8.

Gašević, D, Dawson, S., Rogers, T., & Gasevic, D. (2016) Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success, *The Internet and Higher Education*, 28, 68-84.

Gašević, D. C. Rose, G. Siemens, A. Wolff, and Z. Zdrahal, (2014, March) Learning Analytics and Machine Learning. In *Proceedings Fourth International Conference on Learning Analytics and Knowledge,* Indianapolis, Indiana, USA (pp. 287–288). https://doi.org/10.1145/2567574.2567633

Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision Recall and *F*-Score, with Implication for Evaluation. In D.E.Losada, J.M.Fernández-Luna (Eds.). *Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer*

Science, *3408*. (pp 345-359). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-31865-1_25

Gupta, S., & Sabitha, A. S.(2019). Deciphering the attributes of student retention in massive open online courses using data mining techniques. *Education and Information Technologies*, 24(3) 1973-1994.

Hall, M. A. (1999). Correlation-based feature selection for machine learning. (Ph.D Thesis), University of Waikato, Hamilton, NewZealand.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., ... & Liao, S. N. (2018, July). Predicting academic performance: a systematic literature review. In Proceedings *Companion of the 23rd annual ACM conference on innovation and technology in computer science education*, Larnaca, Cyprus, (pp. 175-199).https://doi.org/10.1145/3293881.3295783.

Hlosta, M., Herrmannova, D., Vachova, L., Kuzilek, J., Zdrahal, Z., & Wolff, A.(2018). Modelling student online behaviour in a virtual learning environment. *arXiv:1811.06369*,[online]Available: http://arxiv.org/abs/1811.06369.

Hong, B., Wei, Z., & Yang, Y. (2017, August 22-25). Discovering learning behavior patterns to predict dropout in MOOC. In *12th International Conference on Computer Science and Education*, Houston, TX, USA (pp. 700-704). doi :10.1109/ICCSE.2017.8085583

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models, *Computer & Education*, *61*, 133-145.

Hung, J.L., Wang, M. C., Wang, S., Abdelrasoul, M., Li, Y., & He, W. (2015). Identifying at-risk students for early interventions—A time-series clustering approach. *IEEE Transactions on Emerging Topics in Computing*, 5(1), 45-55.

Ichiishi, T. (2014). *Game theory for economic analysis*. Elsevier.

Jayaprakash, S.M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, *1*(1), 6-47.

Kim, T., Yang, M., Bae, J., Min, B., Lee, I., & Kim, J. (2017). Escape from infinite freedom: Effects of constraining user freedom on the prevention of dropout in an online learning context. *Computers in Human Behavior*, *66*, 217–231.

Kizilcec, R. F., Pérez-Sanagustín, M., Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses, *Computer & Education*, 104, 18-33.

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014, October 25-29). Predicting MOOC Dropout over Weeks Using Machine Learning Methods. In. Proceedings

Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, (pp. 60–65).

Kőrösi, G., & Farkas, R. (2020 April 24-25). Mooc performance prediction by deep learning from raw clickstream data. *Proceedings in International Conference in Advances in Computing and Data Sciences*. Valletta, Malta, pp 474-485.

Kumar, J.A, & Al-Samarraie, H. (2019). An investigation of novice pre-university students' views towards moocs: The case of malaysia. *The Reference Librarian, 60*(2), 134–147.

Kuzilek, J. Hlosta, M. Herrmannova, D., Zdrahal, Z., & Wolff, A. (2015). Ou analyse: Analysing at-risk students at the open university. *Learning Analytics Review.*, 8, 1-16.

Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). 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. *Computer & Education,72*, 23-36.

Lee, Y., & Choi, J. (2011). A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development*, 59(5), 593-618.

Lim, J. M. (2016). Predicting successful completion using student delay indicators in undergraduate self-paced online courses. *Distance Education*, 37(3), 317-332.

Liyanagunawardena, T. R., Parslow, P., & Williams, S. (2014, February 10-12) . *Dropout: MOOC participant's perspective.* In. *EMOOCs 2014, the Second MOOC European Stakeholders Summit*, Lausanne, Switzerland. (pp 95-100).

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computer & Education, 103*, 1-15.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems, 33*(1), 107-124.

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science*, 8(11), 36.

Nagrecha, S., Dillon, J. Z., & Chawla, N. V. (2017, April). MOOC dropout prediction: lessons learned from making pipelines interpretable. In. Proceedings *26th international conference on world wide web companion*. Perth, Australia. (pp. 351-359). https://doi.org/10.1145/3041021.3054162

Nanopoulos, A., Alcock, R., & Manolopoulos, Y. (2001). Feature-based classification of time-series data. *International Journal of Computer Research*, *10*(3), 49-61.

Palmer,S. (2013). Modelling engineering student academic performance using academic analytics. *International Journal of Engineering Education*, *29*(1), 132-138.

Papamitsiou, Z., & Economides, A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, *17*(4), 49-64.

Park, J., Denaro, K., Rodriguez, F., Smyth, P., & Warschauer, M. (2017, March 21-30). Detecting changes in student behavior from clickstream data. *Proceedings in. 7th International Learning Analytics & Knowledge Conference*, Vancouver, British Columbia, Canada, 21-30. https://doi.org/10.1145/3027385.3027430

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert System with Applications.*, *41*(4), 1432-1462.

Qiu, J., Tang,J.,Liu, T.X., Gong,J., Zhang, C.,Zhang, Q.,& Xue,Y. (2016, February). Modeling and Predicting Learning Behavior in MOOCs. In Proceedings *Ninth ACM International Conference on Web Search and Data Mining*, (pp93-102), New York. https://doi.org/10.1145/2835776.2835842

Ramesh, A., Goldwasser, D., Huang, B., Daum, H., & Getoor, L. (2014). Learning Latent Engagement Patterns of Students in Online Courses. In Proceedings *Twenty-Eighth AAAI Conference on Artificial Intelligence Learning*, *28*(1).

Rolfe, V. (2015). A systematic review of the socio-ethical aspects of massive online open courses. *European Journal od Open, Distance E-Learning,18*(1),52-71.

Saa, A.A. (2016). Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Science and Applications, 7*(5), 212-220.

Salloum S.A., Alshurideh M., Elnagar A., & Shaalan K. (2020) Mining in Educational Data: Review and Future Directions. In. A.E. Hassanien, A. Azar, T. Gaber, D. Oliva, & F. Tolba (Eds.), *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020). AICV 2020. Advances in Intelligent Systems and Computing*, Springer, Cham. https://doi.org/10.1007/978-3-030-44289-7_9

Sha, L., Looi, C.K., Chen, W., & Zhang, B. H. (2012) Understanding mobile learning from the perspective of self-regulated learning. *Journal of Computer Assisted Learning*, 28(4), 366-378.

Shah, D. (2018). By the Numbers: MOOCS in 2018 Class Central (2018). Retrieved December 16, 2018, from https://www.classcentral.com/report/mooc-stats-2018/.

Sharkey, M., & Sanders, R. (2014, October 25-29). A process for predicting MOOC attrition. In *Proceedings* e 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, (pp. 50-54).

Soffer T., & Cohen, A. (2019). Student's engagement characteristics predict success and completion of online courses. *Journal of Computer Assisted Learning*, 35(3), 378-389.

Stillson, H., & Alsup, J. (2003). Smart ALEKS... or not? Teaching basic algebra using an online interactive learning system. *Mathematics and Computer Education*, *37*(3), 329.

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behaviour*, 47, 157-167.

Wang, J., Xu, M., Wang, H., & Zhang, J. (2006, November). Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. In *2006 8th international Conference on Signal Processing* (Vol. 3). IEEE.

West, D. M. (2012). Big Data for Education: Data Mining, Data Analytics, and Web Dashboards, Gov. Stud. Brookings, US Reuters, no. 1.

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist*, 45(4), 267-276.

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013, April 8-12). Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. Proceedings *In: Third Conference on Learning Analytics and Knowledge,* Leuven, Belgium.

Zacharis, N. Z. (2015) A multivariate approach to predicting student outcomes in Web-enabled blended learning courses. *The Internet and Higher Education*, 27, 44-53.