Machine Learning Classification Algorithms for Systematic Analysis to Understand Learners Drop out of MOOCs courses

Seema Rawat
Amity University
Deepak Kumar (deepakdeo2003@gmail.com)
Amity University https://orcid.org/0000-0003-4487-7755
Chhaya Khattri
Amity University
Praveen Kumar
Amity University

Research Article

Keywords: Educational Data Mining, MOOC, Student Dropout, Machine Learning, Prediction

Posted Date: May 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-491528/v1

Abstract

The increasing popularity of massively online open courses (MOOCs) has been attracting a lot of learners. Despite the popularity, it has been observed that there is a significant percentage of learners who discontinue courses and drop out of the platform. This is a problem that most of the MOOC courses face. The dropout probability of any student depends on his/her interaction with the platform, and the features of the course in which the student has enrolled. The research work is intended to study and analyze the dropout behavior of the students in online learning with identification of the reasons and to understand their impact. The current research accounts for the activity log of learners of 13 different online courses offered by Harvard and MIT during 2012 to 2013. The work examines the attributes which affects the student dropout rate. The research can be useful in improving the existing features of the MOOC courses and content to ensure persistence turnout of their learners.

1. Introduction

The prevalence of MOOCs at present is quite at its peak. The MOOC serves as a good alternative to geographical incompatibilities, or when a student cannot manage attending regular classes. Since MOOCs are easily accessible from anywhere and anytime, they naturally become a preferred choice. Most MOOCs are self-paced, so that the student can study comfortably. Coursera leads the MOOC industry with over 53 million students and thousands of specialization courses and degree programs. Followed by edX and Udacity, there isn't any subject matter, theme or field rendered untouched by these platforms. MOOCs provide their students an opportunity to harness the knowledge at ease. According to Class Central's MOOC report (edX, 2018), it has been observed that most of the online learners are beyond the age of 25, that categorizes them as 'continuing learners', suggesting that either they take up these courses in addition to their full time studies, or due to the need of updating oneself as required by the professional environment. The onset of successful MOOCs was marked by an online course of 'Artificial Intelligence' in 2011 fall. It was organized by ex-Stanford professors Sebastian Thrun and Peter Norvig (who founded Udacity the same year). There was a huge number of enrollments from the world over, and thus, by 2012, the era of online learning through MOOCs set in.

Even though the introduction of MOOCs as a part of (distant) learning has transformed the teaching system and benefitted many individuals, there are persistent problems that most MOOCs still face after over eight years of their foundation, that is, a high dropout/withdrawal rate of students. A research by Justin et al. (2019) suggests that the position of low completion rate of MOOCs has not been improved since the past six years. Nearly 52% of students never explored the course content after their enrollment, and the dropout rate is highest in the first two weeks of course initiation.

This fact is confirmed by Rosé, C. P. et al. where it has been asserted that it is important to keep the student interested and engaged during the first two weeks to avoid attrition. Also, a research by Jacobsen, D. Y. (2019) claims that the learners who explore less pages of the course, and do not attempt and submit assignments regularly, are more likely to dropout; however, if the learners apply whatever knowledge that

is gained through a MOOC, in practicality, then it is a positive outcome of the course. To keep the record of the students' intention and understand them better, HarvardX conducts a pre-course survey that includes questions not only about general information of the student, but also about the reason of enrollment, confidence regarding completion of the course, and most importantly, asks the student to formulate a goal plan on how is he/she going to accomplish his goal. HarvardX has taken a step towards the examination of the reasons why students may tend to drop out, and that is suggested by/asked to the enrollees themselves. Figure 1 shows a fragment of a Harvard University course survey form.

This research, it is intended to observe the trends in the characteristics of a typical dropout student, and also study the gravity of different attributes or factors resulting in the same. There are many purposes for which the students enroll in a course. It can be for exploring the course structure, aspiring to earn a certificate or just out of curiosity. Monitoring the clickstream activity and performance of a learner, it is possible to identify the probability of a student completing a particular course. The insights can be achieved by applying educational data mining upon the student data. The rationale behind this research is to bring out the underlying factors affecting attrition and dropout rate of students to light.

2. Literature Review

There can be numerous reasons that result in people dropping out, or people completing a course but never coming back to a particular website again. Now that the dropout is a popular interest among researchers, there is a huge amount of relevant literature available that focus on this particular area. Table 1 gives an overview of the authors and the techniques that they have used to address the problem to give possible solutions. It has been observed that most of these research works are based on clickstream analysis and study of patterns in student engagement.

Table 1
Summary of Significant Work in Dropout Prediction for MOOC

Author(s)	Year	Research Question	Methodologies/Tools	Findings
Anat Cohen and Orit Baruth	2017	Q1: What are the personal characteristics of the online course students? Q2: What is the level of satisfaction of the online students? Q3: Is there a correlation between personality traits and satisfaction with online courses? Q4: Can students' satisfaction from online courses be predicted by their personality traits? Q5: Is it possible to characterize groups of online learners according to their personality and satisfaction with the online course? Q6: Are there differences among groups relating to synchronous or asynchronous learning?	Questionnaires, characterization of students	It is important to develop course formats that enable the use of SRL (Self-regulated Learning) strategies for those who are Characterized with a high-conscientiousness personality and openness to experience. However, these course formats should also be suitable to those learners who do not tend to high conscientiousness or openness to experience, and, therefore, may exhibit a lower degree of SRL strategies, or none at all.

Author(s)	Year	Research Question	Methodologies/Tools	Findings
Dr. Marcela Georgina Gomez-Zermeño and Lorena Alemán de la Garza	2016	To identify the terminal efficiency of the Massive Online Open Course “Educational Innovation with Open Resources” offered by a Mexican private university.	Statistics and Probabilistic models, surveys	Those who decided to leave the course indicated problems with the structure and guidance in the course, limitations on the use of information technology or in English, in addition to the limited availability of time due to family or work reasons. The probability of abandonment decreases when participants are over 55, have a strong commitment to the MOOC and when they have full or partial employment.
Boyer and Veeramachaneni	2015	Does transfer learning provide accurate insights and prediction results regarding MOOC stop out?	Logistic Regression, Multitask learning, Transfer Learning	These research results are a foundation for developing more robust and advances predictive models to analyses student engagement in MOOCs.

Author(s)	Year	Research Question	Methodologies/Tools	Findings
Coleman et al.	2015	Can LDA serve as an unsupervised approach for discovering the behavioral trends of MOOC participants? Can the mixed-membership model from LDA predict certification?	NLP, LDA	The course content and interaction with the courseware is taken into consideration, and it suggested that the predictive results highly depend on the students' use case proportion and the total number of certificate earners and non-earners
Kizilcec et al.	2015	How are learners' characteristics related to their persistence and performance in the course? How is learners' satisfaction related to their persistence in the course? What reasons do learners report for disengaging from a course? How do reasons for disengaging vary by learner characteristics?	Surveys, Logistic mixed effect model	84% of the learners cannot complete/drop out of a course because they could not make enough time for it.
Taylor et al.	2014	Is it possible for machine learning algorithms, with only a few weeks of data, to accurately predict persistence? Is it possible to predict, given only the first week of course data, who will complete the last week of the course? How much history (or how many weeks' data) is necessary for accurate prediction one or more week ahead?	Logistic Regression, SVM, Deep belief Networks, Decision Trees	Features which incorporate student problem submission engagement are the most predictive of stopout.

Author(s)	Year	Research Question	Methodologies/Tools	Findings
Miaomiao Wen et al.	2014	How do student opinions affect MOOC attrition rate?	Sentiment Analysis	Students who are exposed to a standard deviation more negativity are 6% more likely to drop out on the next time point than students who are exposed to some amount of positivity.
Rebecca M Stein and Gloria Allione	2014	To analysis attrition in a particular MOOC and extract the impacting features.	Surveys, Cox Proportional Hazard Model	Students with MOOC platform experience are approximately 20 percent less likely to drop out. Students participating in course quizzes and per-assignments are about 65% less likely to dropout.
Liyanagunawardena, T. et al.	2014	To collect and present personal opinions of MOOC students on MOOC attrition and success rates.	Interviews	For MOOC participants, ‘dropout’ means achieving their aims (or not) in a course rather than finishing the course by completing all parts.
D.F.O.Onah et al.	2014	What aspects of learner behavior affect the high dropout rates in MOOCs the most?	Comparative Case Study	A lack of in-person support affects student engagement, and students learn better when allowed to study at their own pace.
Yang, D. et al.	2013	How to create a social environment that would be more conducive to promoting continued engagement in MOOCs?	Survival Analysis, social network analysis	Students actively participating in forums and text discussions in the first week are 33% less likely to dropout of a MOOC.

Author(s)	Year	Research Question	Methodologies/Tools	Findings
Ramesh, A. et al.	2013	What are the key factors influencing learner performance in an online setting?	Probabilistic Soft Logic	A student's passive engagement had a high weight in prediction, and engaged learners are likely to post content with negative sentiment on the course, which may be perceived as participating in the course.
Balakrishnan & Coetzee	2013	Can we accurately predict whether a student is likely to drop the MOOC in the near future? Can we identify patterns in the behavior of students who eventually drop the course, and thus suggest interventions to prevent this from occurring?	HMM and SVM	Students who rarely or never post on forums, and who rarely or never check their progress drop the most (37–40%).

The most common approaches among the ones discussed above are logistic regression and survival analysis. Based on LR, the authors like Boyer, S., & Veeramachaneni, K. (2015), Taylor, C. et al. (2014), He, J. et al. (2015) and Kizilcec, R. F., & Halawa, S. (2015) have proposed dropout prediction models. Also, Yang, D. et al. (2013) and others have deployed survival analysis to examine the impact of various factors, which present the likelihood of a failure in a point of time. Liyanagunawardena, T. et al. (2014), Cohen, A., & Baruth, O. (2017), D.F.O. Onah et al. (2014) and Gomez-Zermeno, M. G., & Aleman De la Garza, L. (2016) used unique approach to incorporate behavioural analysis to understand the student engagement and its impact on their learning process, whereas Liu, T. et al. used clustering to obtain various insights of the pool of students and thus they differentiated the students' group on the basis of activity.

Ramesh, A. et al. (2013) used probabilistic soft logic to analyse the weights of different drop out factors. Balakrishnan & Coetzee (2013) had used Hidden Markov Model and Support Vector Machines for a better prediction model creation. "HMMs associate different states of the data with a probability distribution. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes very well." Lastly,

Rebecca M Stein and Gloria Allione (2014) also used the Cox proportional hazard model “which assumes that the covariates affect dropout in a proportional manner that is time invariant”.

Systematic Literature Review:

We have observed the related work to get an overview of current status of the research related to machine learning classification techniques. As the field of machine learning classification is developing rapidly (e.g., application in academic sector with deep learning) to provide an outline of the current trends of these procedures. Moreover, we see that most other secondary studies emphasis on the study of different techniques as a part of the systematic literature review process. This study aims to offer an over-view of all steps of the process. The current work analyses the following keywords for the analysis “*Machine”, “Learning”, “Classification”, and “MOOC”*.

Analysis of Keywords: “*Machine*” + “*Learning*” + “*Classification*” + “*MOOC*”

The above Fig. 2 represents the year wise publication volume of documents with “*Machine*” + “*Learning*” + “*Classification*” + “*MOOC*” keywords over the years 2013–2021. Volume of documents publication has raised through 2014 onwards.

The above Fig. 3 characterizes the year wise documents type over the years 2013–2021. It can be observed that year wise documents type has elevated over 2016 onwards.

The above Fig. 4 signifies the year wise documents published by different authors over the years 2013–2021 having publication count up to 15 publications. It can be observed that the maximum publication of documents by individual author is 10.

The above Fig. 5 characterizes the documents published by different country over the years 2013–2021. It can be observed that the China has published more number of documents with 225 in number followed by United States with 200 in number.

The above Fig. 6 describes the documents published by different country over the years 2013–2021. It can be observed that the maximum published document is in the form of conference paper with 44.6% followed by review document with 3.5%.

The above Fig. 7 describes the documents published in different subject over the years 2013–2021. It can be observed that the maximum published document is engineering domain with 12.6 % followed by mathematics with 8.7% and decision sciences with 4.4%.

The above Fig. 8 describes the different funders over the years 2013–2021. It can be observed that the maximum research is being funded by the National Natural Science Foundation of China with 90 + documents followed by National Science Foundation with 50 and European Commission with 35

From Table 1, and referring to various other research, the commonly used attributes for analysis were identified. These have been outlined in Fig. 9.

It can be inferred from Table 2 that the most researched attributes in existing literature are student engagement, student feedback & reviews and also, the quality and the type of content the course offers. It can be seen that the language barrier, personality of students (punctuality) and freedom of choice of course (skipping assignments, tutorials) has not been the popular choice, hence, some of these factors are also considered in this research. The considered dataset (Refer Table 3) includes some of the new data introduced for evaluation purpose.

2.1 Description of Machine Learning Algorithms used

2.1.1 Decision Tree

The Decision Tree is a supervised learning technique that branches a dataset on the basis of different features. From the head node to the root node, the tree will pick the most important input attribute and split branches according to the class label predictions. The end node will always be the column values needed to be predicted. Hence, the appropriate prediction according to the inputs, will be the end node in the tree.

2.1.2 Logistic Regression-

Logistic Regression is a supervised classification algorithm used to predict binary dependent variables, resulting into a figure of probability that strictly lies between 0 and 1. In this case, the dependent variable is the column having values of the students being a drop out or not. The algorithm presents P(Y = 1) as a function of X.

2.1.3 Neural Networks-

The Neural Networks are a set of algorithms inspired by human neurons, (that are modeled as perceptrons in machine learning) that are used to identify pattern in large data. They label, cluster and classify raw input data to identify the pattern, and arrange them into layers of class-wise input nodes, thereby containing the pattern in vectors.

2.1.4 Gradient Boosted Trees-

The Gradient Boosting algorithm means turning weak learners in a training dataset into strong learners, usually by the means of decision tree. Modifying the weights of features that are difficult to classify, and using those as a base for the next improved trees, the predictions of the final tree is a summation of the previous boosted trees.

2.1.5 Deep Learning-

The Deep Learning is an implementation of artificial neural networks. The algorithm processes data in the form of several layers (more layers than a neural network, hence the name ‘deep’), to draw conclusion based on it. It uses different input data to produce insights and then draw larger conclusions (or predictions). This is similar to human’s behavior of ‘learning by experience/example’.

3. Methodology

3.1 Dataset selection

With reference to the attributes in Table 2, the chosen dataset is from Harvard and MIT MOOCs that includes relevant information. A detailed description about the dataset is given in Sect. 4.1.

3.2 Data preprocessing

To make the data useful, all the missing, invalid values were eliminated. Furthermore, to improve the integrity of the data, the independent variables were normalised using ‘Nominal to Numerical’ operator in RapidMiner Studio. This helped in acquiring fast and precise result because all the values follow the same scale. Using ‘Split Data’ operator, the dataset was divided into two parts: 75% for the training, and 25% for the prediction test.

3.3 Feature Selection

The prepared dataset was examined for the most relevant input feature. A research by Gupta, S., & Sabitha, A. S. (2019) presented its findings according to the clickstream analysis of the students and discussion participation, that included ‘number of times videos played’, ‘exploration of course’ and ‘number of forum posts by student’. In this research, four new features were introduced in the dataset. Therefore, a total of 12 attributes were chosen as input for the prediction model. The dataset fields are explained in Table 3 of Sect. 4.

3.4 Determining the significance of given characteristics

Five different machine learning algorithms were used in this research so as to compare the weights of the input features, to determine the important ones that impact the dropout rate of students in any MOOC.

4. Experimental Setup

4.1 Dataset

The data used for analysis is from the 2012–2013 academic year of the first year of introduction of MOOC by HarvardX and MITx, retrieved from Kaggle. The features of this dataset are explained in Table 2.

Table 2: Metadata of Dataset

Attribute	*Type*	*Description/Source*
'Course_id'	Varchar	The unique ID for the course a student is enrolled in.
'year'	Year	The academic year for the course.
'Semester'	String	The semester of the particular course (Fall or Spring).
'Userid_DI'	varchar	The unique ID of a student.
'registered'	binary	Holds whether a student has registered for a course (0) or not (1).
'Viewed'	binary	Holds whether a student has viewed the course (0) or not (1).
'explored'	binary	Holds whether a student has explored related courses (0) or not (1).
'certified'	binary	Holds whether a student is certified (0) or not (1).
'final_cc_cname_DI'	String	The country that the student belongs to.
'LoE_DI'	String	The highest level of education completed by the student.
'gender'	String	The gender of the student.
'Grade'	double	The grade obtained by a student in a course.
'start_time_DI'	date	The date of the first recorded course activity of the student.
'last_event_DI'	date	The date of the last recorded course activity of the student.
'Nevents'	integer	The number of course events the student has participated in.
'ndays_act'	integer	The number of course events the student has participated in.
'nplay_video'	integer	The number of times the student has played course videos.
'Nchapters'	integer	The number of course units the student has completed.
'nforum_posts'	integer	The number of forum posts the student has made.
'incomplete_flag'	binary	Holds whether a student has completed a course (0) or not (1).
'Age'	integer	The age of the student.
'duration'	integer	The duration of a particular course in days.
'date'	date	The start date of each of the courses offered. (via Ho, A. et al.)
'difference'	integer	The difference between 'start_act' and 'date', i.e., the difference in the initial activity of the student (in days).
'Lang'	String	The English Proficiency Index of a country (2012-2013), as reported by EPI.

*These columns and their data entries are not present in the original dataset. These data were collected using internet sources.

4.2 Tool used

To achieve desired results, RapidMiner Studio v9.6 has been used. According to Edutechwiki, RapidMiner Studio is a "downloadable GUI for machine learning, data mining, text mining, predictive analytics and business analytics". It helped to identify weights and observe patterns of the different attributes under consideration that make an impact on student attrition.}

5. Case Study 1

Pilot Study of the MOOC Dataset

5.1 Demographics

A brief study of the demographics of the considered dataset in presented below. It was observed that out of 367,375 participants, 73.6% were male and 26.4% were female. Also, the majority held the maximum education level as Bachelor's while as many as 47.2% belonged to an area with language proficiency as 'high'. Nearly 81% of the participants belonged to the age group 15–32, while 14.65% were between 33 and 49. A small fraction of 0.35% was over 65. These courses had the most participants from the United States and India, followed by Europe, Africa, United Kingdom and Brazil. Some of the popular courses observed were Introduction to Computer Science and Programming (6.00x- Fall and Spring), Circuits and Electronics (6.002x- Fall), Health in Numbers: Quantitative Methods in Clinical & Public Health Research (PH207x- Fall) and Justice (ER22x- Spring).

5.2 Characteristics of dropouts evident from the dataset

The distribution of the noticeable characteristics of the dropout students can be explained as follows. Nearly 3.128% of the 2012–2013 batch of Havard/MIT MOOC learners did not complete their respective courses. It was observed that the majority of the students had dropped out of three courses, 6.00x (by Harvard), 8.02x (Electricity and Magnetism) and 14.73x (Challenges of Global Poverty) (by MIT). The students who dropped out of their respective courses held bachelor's level of education. The US, India and Europe had the greatest number of dropouts. The least number of dropouts were from East Asian countries.

The statistical study of this dataset revealed that the dropouts mainly belonged to a certain age group and country. This confirms the impact of demographics on the retention of a student. The maximum number of dropouts were from one particular course (6.00x), which implies that the course in which a student is enrolled in, may affect his/her engagement. It was also seen that the dropouts here held bachelor's level of education, and the ones with the highest education level as 'doctorate' comprised the least of the dropouts (2.7%). This implies that the education level of a student must match the prerequisites of the course taken (for example, 6.00x was an undergraduate level course, and 32.3% of the dropouts had highest education as secondary or less).

However, this study did not focus on all course-related and student-related factors that were derived through the literature survey. For this purpose, the new data according to those features were gathered through appropriate resources, to be examined as well (Refer Sect. 3.2). These data were then incorporated into the research methods adopted, to retrieve new relevant results.

6. Case Study 2

Prediction of Student Attrition

The attributes chosen for the further analysis of the dataset to produce insights about dropout rates were- 'date', 'difference', 'duration', 'lang', 'course_id', 'semester', 'final_cc_cname_DI', 'LoE_DI' (Loe),'Gender', 'Age', 'start_time_DI', and 'incomplete_flag'. To understand the impact of these new features, five predictive machine learning algorithms: deep learning, neural networks, decision tree, logistic regression and gradient boosted trees, were applied to the inputs, so as to identify the weights of all the features, and determine if they are some new dominant features that can contribute towards the tendency of a student dropping out of an online course. The machine learning models predicted whether a student will drop out (1) or not (0). The weights of the features thus calculated, suggested their respective impacts on the prediction results. The following section describes these results.

6.1 Attribute Weights

The student attrition probability was predicted using various input values so as to determine which attribute had the most impact on the results. Each algorithm assigned weights to the input features according to the prediction results. A brief description of the same is given in the subsequent paragraphs:

i) Neural Network- The algorithm laid emphasis on three input attributes, i.e., 'semester = Spring', the weight by correlation being 0.138; 'semester = Fall'- the weight by correlation being 0.133 and 'course_id = 8.02x'- the weight by correlation being 0.108, which roughly calculates to 13.8%, 13.3% and 10.8% Figure 5 shows the values of Sigmoid function for an improved neural network formation.

ii) Decision Tree- According to this algorithm, the attributes having the highest weights were 'difference'- with a weight of 0.416, and 'age'- with a weight of 0.The decision tree for this prediction is given in Fig. 6 below. It shows that the first attribute for branching is 'difference', because it has the greatest impact on the result.

iii) Logistic Regression- Using this algorithm, it was observed that 'course_id' being 6.00x, 6.002x or CB22x meant a greater probability of a student dropping out of these This implies that the course contents play an important role in dropout behavior of a The combined weight of these features was calculated as 14.825%. It was seen that students who belonged to the United States or India, were more likely to drop Also, attributes like 'semester' as Fall weighted in the negative scale, and 'lang' (language proficiency of the carried a near-zero A bar chart comparing the weights of various attribute values in given in Fig. 7.

iv) Gradient Boosted Trees- The algorithm created 100 progressive decision trees, portraying the best possible attribute relations so that their weights could be The results revealed that 'difference' feature was present in most of the dropout cases (over 75,000 students). The most dominant attribute here was 'duration', with a weight of 19258.It was also seen that the 'course_id = 6.002x' had an impact on the Again, this clearly lays emphasis on the duration, along with the type of course content the student is Figure 8 displays a bar chart for the same.

v) Deep Learning- Through the implementation of a deep learning prediction model, it was observed that the top three highest attribute weights were 'semester = Spring', 'semester = Fall' and 'course_id = 8.002x', constituting 37.9% of the feature This implied that the course content and its time of offering are important in determining whether a student drops out or Figure 9 shows a pictorial comparison of the input features.

6.2 Performance comparison

6.2.1 Accuracy and Execution Time

Table 3
Performance of Algorithms Used

Algorithm	Accuracy (%)	Execution Time
Neural Network	96.97	44m 49s
Decision Tree	96.88	1m 24s
Logistic Regression	96.87	4s
Gradient Boosted Trees	96.25	4m 42s
Deep Learning	89.06	1m 31s

According to the respective performance matrices of the algorithms, 4 out of 5 algorithms indicated accuracy above 95%. Out of the five algorithms used, Neural Networks had the highest accuracy (96.97%), and Deep Learning had the lowest (89.06%). The Logistic Regression executed the fastest with an execution time of 4 seconds, and the slowest algorithm was Neural Network, with an execution time of 44 minutes and 49 seconds. It can be thus inferred that the Decision Tree and the Deep Learning were the most optimal algorithms. Table 3 summarizes the different accuracy levels and execution time exhibited by all the algorithms

6.3 Summary

It was confirmed, that the features and attributes that were considered by the authors in previous literature, the highest number of times (Refer Table 3), i.e., 'student engagement', 'quality and type of content', 'student feedbacks' and 'demographics', have a huge impact on a MOOC learner's tendency to drop out of the course. Four out of five algorithms emphasized on 'course_id', which equals to the type of

course and its contents. In addition to these, three more attributes came out dominant in this area. Deep Learning, Gradient Boosted Trees and Neural Networks, suggested ‘semester’ (the time of year a particular course is offered) as an important attribute in affecting the dropout rate, whereas according to the Decision Tree, and the Gradient Boosted Trees, ‘difference’ (the difference in days between the student’s first activity and the course’s start date) was a salient feature. It is important to note that Gradient Boosted Trees result also suggested the influence in the dropout behavior depending on the duration of a course. Table 4 below summarizes the prediction results related to the attribute weights.

Table 4: Important Factors Affecting Dropout in a MOOC

Algorithm Used for Prediction	Important Factors Affecting Dropout in a MOOC					
	Course Content	Semester of Course	Birth Country (Student)	Age of Student	Activity Difference	Duration of Course
Deep Learning	✓	✓				
Gradient Boosted Trees	✓	✓			✓	✓
Logistic Regression	✓		✓			
Neural Network	✓	✓				
Decision Tree				✓	✓	

7. Conclusion

This paper advances research work to study and analyze the dropout behavior of the students in online learning with identification of the reasons and to understand their impact.

Our systematic literature review identifies the theoretical approaches which are most used to study the development, potential and dynamics. The work examines the attributes which affects the student dropout rate. The research can be useful in improving the existing features of the MOOC courses and content to ensure persistence turnout of their learners. It was made clear through this research that the likelihood of a student dropping out of a MOOC essentially depends on both course- related and student’s behavioural factors. As an outcome of the previous literature analysis, many factors were marked as important in making an impact in this area. There was a great emphasis on the clickstream pattern of a student, and interaction with the co-learners, as this is the most obvious way to infer the interest of a student in a course. The dropout rate in the student data of 13 MOOCs and over 300,000 learners was calculated using 5 machine learning (predictive) models: Deep Learning, Gradient Boosted Trees, Logistic Regression, Decision Tree and Neural Networks. The performance matrix of these algorithms displayed class prediction accuracy above 88%. Most of the dropout students belonged to the age group of 19 to 25, and they were majorly from United States, India and Europe. The results revealed that the significance

of a few less explored features not addressed in the previous works. For example, the highest number of dropouts were from the course ‘6.00x’, which was ‘Introduction to Computer Science and Programming’. It was also seen that if a student had a first activity difference of more than 20 days, he is more likely to drop out of the respective course. The results are consistent with Banerjee and Duflo (2014), who found that students who enroll one day late are 17 percent less likely to earn a certificate than students who enroll on time, confirming that patterns of retention can be detected by early behavior. Hence, it was deduced that, in addition to the student engagement intensity, the course content, the time of the course offering, and punctuality of a student regarding the course activities, and sometimes, the duration of the course also made a crucial impact in the dropout trend in MOOCs.

8. Declarations

Author Contributions

Dr Seema Rawat conceived and designed the study, Ms Chhaya Khattri performed the research, Dr Deepak Kumar analyzed the data, and Dr Praveen Kumar contributed to editorial input.

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

9. References

1. Allione G, Stein RM (2016) Mass attrition: An analysis of drop out from principles of microeconomics MOOC. J Econ Educ 47(2):174–186
2. Balakrishnan G, Coetzee D (2013) Predicting student retention in massive open online courses using hidden markov models. Electrical Engineering Computer Sciences University of California at Berkeley 53:57–58
3. Banerjee AV, Duflo E (2014) (Dis) organization and success in an economics MOOC. Am Econ Rev 104(5):514–518
4. Boyer S, Veeramachaneni K (2015, June) Transfer learning for predictive models in massive open online courses. In *International conference on artificial intelligence in education* (pp. 54–63). Springer, Cham
5. Brunskill E, Zimmaro D, Thille C (2018, June) Exploring the impact of the default option on student engagement and performance in a statistics MOOC. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (pp. 1–4)
6. Cohen A, Baruth O (2017) Personality, learning, and satisfaction in fully online academic courses. Comput Hum Behav 72:1–12
7. Coleman CA, Seaton DT, Chuang I (2015, March) Probabilistic use cases: Discovering behavioral patterns for predicting certification. In *Proceedings of the Second (2015) ACM Conference on*

Learning@ Scale (pp. 141–148)

8. Feng W, Tang J, Liu TX (2019, July) Understanding dropouts in MOOCs. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 517–524)
9. Gomez-Zermeno MG, Aleman De la Garza, L (2016) Research Analysis on MOOC Course Dropout and Retention Rates. Turkish Online Journal of Distance Education 17(2):3–14
10. Gupta S, Sabitha AS (2019) Deciphering the attributes of student retention in massive open online courses using data mining techniques. Education Information Technologies 24(3):1973–1994
11. He J, Bailey J, Rubinstein BI, Zhang R (2015, February) Identifying at-risk students in massive open online courses. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*
12. Ho A, Reich J, Nesterko S, Seaton D, Mullaney T, Waldo J, Chuang I (2014) HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013. *Ho, AD, Reich, J., Nesterko, S., Seaton, DT, Mullaney, T., Waldo, J., & Chuang, I.(2014). HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No. 1)*
13. Jacobsen DY (2019) Dropping out or dropping in? A connectivist approach to understanding participants' strategies in an e-learning MOOC pilot. Technology Knowledge Learning 24(1):1–21
14. Kizilcec RF, Halawa S (2015, March) Attrition and achievement gaps in online learning. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 57–66)
15. LIU TY, Xiu LI (2017) Finding out reasons for low completion in MOOC environment: an explicable approach using hybrid data mining methods. *DEStech Transactions on Social Science, Education and Human Science*, (meit)
16. Liyanagunawardena TR, Lundqvist K, Williams SA (2015) Who are with us: MOOC learners on a FutureLearn course. Br J Edu Technol 46(3):557–569. doi:10.1111/bjet.12261
17. Onah DF, Sinclair J, Boyatt R, Foss J (2014, November) Massive open online courses: learner participation. In *Proceeding of the 7th International Conference of Education, Research and Innovation* (pp. 2348–2356)
18. Ramesh A, Goldwasser D, Huang B, Daumé III, H., & Getoor L (2013, December) Modeling learner engagement in MOOCs using probabilistic soft logic. In *NIPS workshop on data driven education* (Vol. 21, p. 62)
19. Reich J, Ruipérez-Valiente JA (2019) The MOOC pivot. Science 363:130–131. 10.1126/science.aav7958
20. Rosé CP, Carlson R, Yang D, Wen M, Resnick L, Goldman P, Sherer J (2014, March) Social factors that contribute to attrition in MOOCs. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 197–198)
21. Sinha T (2014) Who negatively influences me? Formalizing diffusion dynamics of negative exposure leading to student attrition in MOOCs. *arXiv preprint arXiv:1407.7133*
22. Sinha T, Li N, Jermann P, Dillenbourg P (2014) Capturing" attrition intensifying" structural traits from didactic interaction sequences of MOOC learners. *arXiv preprint arXiv:1409.5887*

23. Taylor C, Veeramachaneni K, O'Reilly UM (2014) Likely to stop? predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382*
24. Wang YC, Kraut R, Levine JM (2012, February) To stay or leave? The relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 833–842)
25. Wen M, Yang D, Rose C (2014, July) Sentiment Analysis in MOOC Discussion Forums: What does it tell us?. In *Educational data mining 2014*
26. Yang D, Sinha T, Adamson D, Rosé CP (2013, December) Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop* (Vol. 11, p. 14)
27. Yang D, Wen M, Rose C (2014, July) Peer influence on attrition in massively open online courses. In *Educational data mining 2014*
28. Cui Y, Jin WQ, Wise AF (2017) Humans and machines together: Improving characterization of large scale online discussions through dynamic interrelated post and thread categorization (DIPTiC). Paper presented at the *L@S 2017 - Proceedings of the 4th (2017) ACM Conference on Learning at Scale*, 217–219. doi:10.1145/3051457.3053989 Retrieved from www.scopus.com
29. Cui Y, Wise AF (2015) Identifying content-related threads in MOOC discussion forums. Paper presented at the *L@S 2015–2nd ACM Conference on Learning at Scale*, 299–303. doi:10.1145/2724660.2728679 Retrieved from www.scopus.com
30. Cui Y, Wise AF, Allen KL (2019) Developing reflection analytics for health professions education: A multi-dimensional framework to align critical concepts with data features. Comput Hum Behav. doi:10.1016/j.chb.2019.02.019
31. Wise AF (2018) Learning analytics: Using data-informed decision-making to improve teaching and learning. *Contemporary technologies in education: Maximizing student engagement, motivation, and learning* (pp. 119–143) doi:10.1007/978-3-319-89680-9_7 Retrieved from www.scopus.com
32. Wise AF, Cui Y (2018) Envisioning a learning analytics for the learning sciences. *Proceedings of International Conference of the Learning Sciences, ICLS, 3*(2018-June), 1799–1806. Retrieved from www.scopus.com
33. Wise AF, Cui Y (2018) Learning communities in the crowd: Characteristics of content related interactions and social relationships in MOOC discussion forums. Computers Education 122:221–242. doi:10.1016/j.compedu.2018.03.021
34. Wise AF, Cui Y (2018) Unpacking the relationship between discussion forum participation and learning in MOOCs: Content is key. Paper presented at the *ACM International Conference Proceeding Series*, 330–339. doi:10.1145/3170358.3170403 Retrieved from www.scopus.com
35. Wise AF, Cui Y, Jin WQ (2017) Honing in on social learning networks in MOOC forums: Examining critical network definition decisions. Paper presented at the *ACM International Conference Proceeding Series*, 383–392. doi:10.1145/3027385.3027446 Retrieved from www.scopus.com

36. Wise AF, Cui Y, Vytasek J (2016) Bringing order to chaos in MOOC discussion forums with content-related thread identification. Paper presented at the *ACM International Conference Proceeding Series„ 25-29-April-2016* 188–197. doi:10.1145/2883851.2883916 Retrieved from www.scopus.com
37. Wise AF, Schwarz BB (2017) Visions of CSCL: Eight provocations for the future of the field. International Journal of Computer-Supported Collaborative Learning 12(4):423–467. doi:10.1007/s11412-017-9267-5
38. http://edutechwiki.unige.ch/en/RapidMiner_Studio
39. https://www.classcentral.com/report/edx-2018-review/
40. https://www.ef.com/wwen/epi/
41. https://www.kaggle.com/kanikanarang94/mooc-dataset

Figures

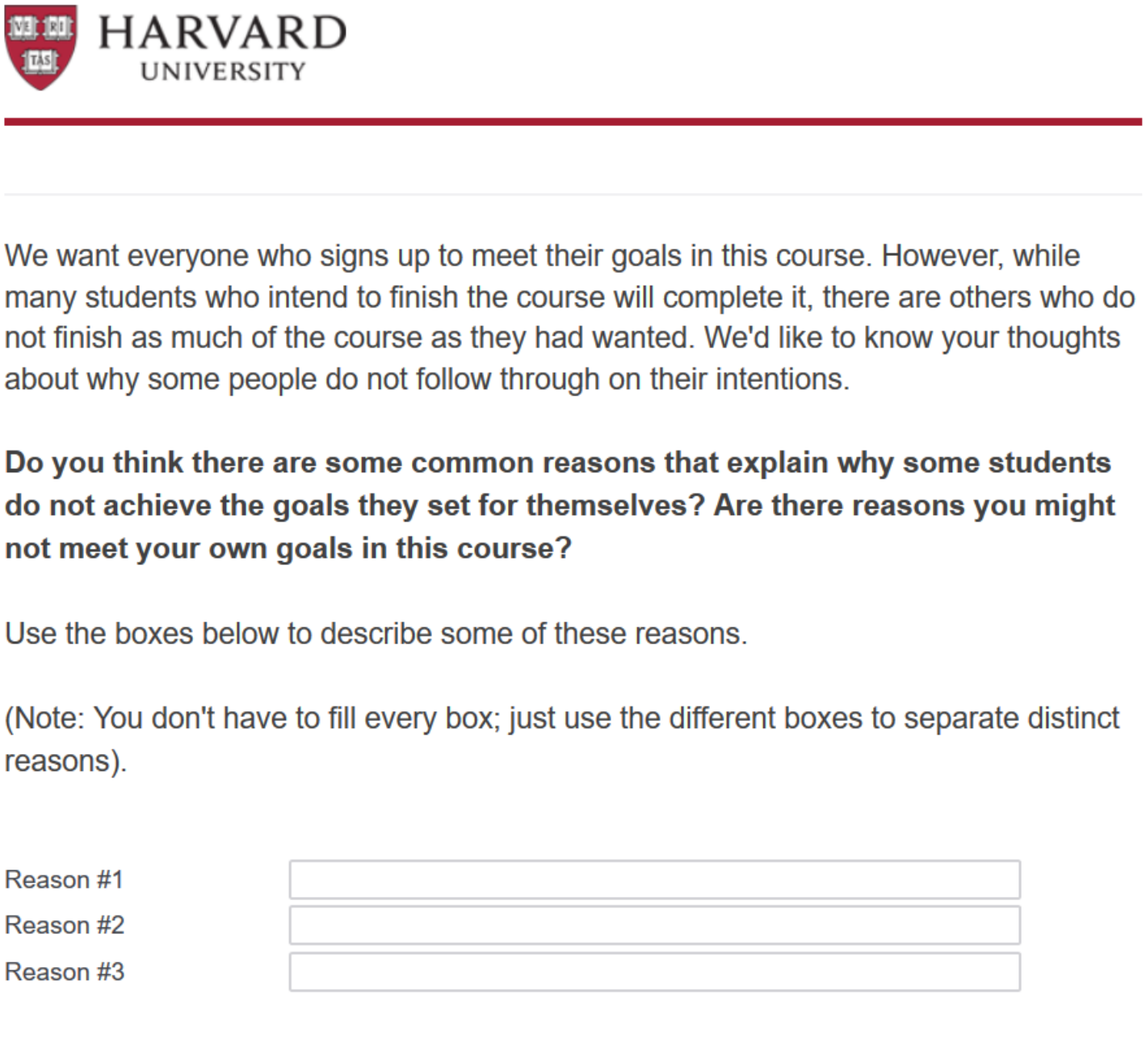

Figure 1

Pre-course Survey for HarvardX/PH526x

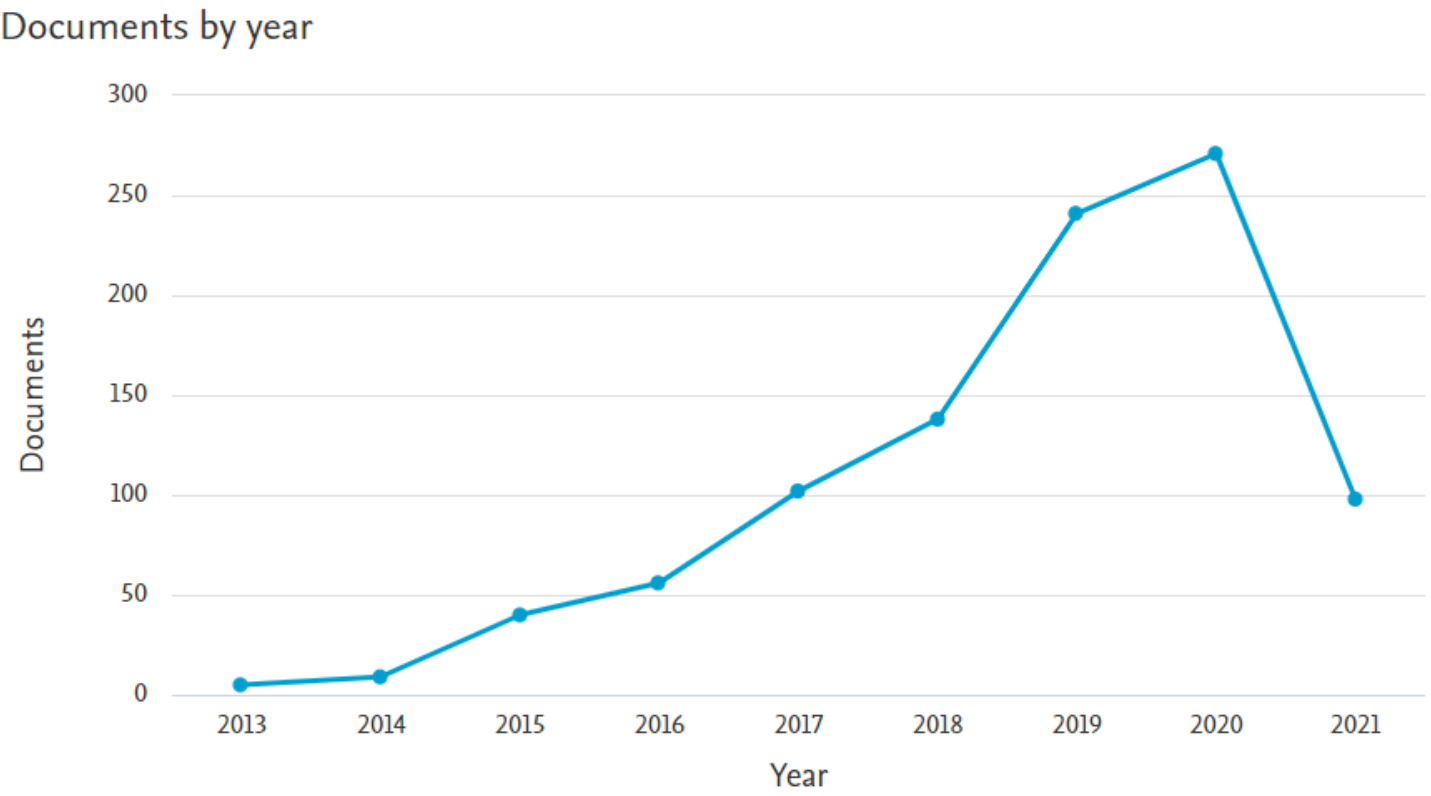

Figure 2

Survey of year-wise publication of documents over the years 2013-2021

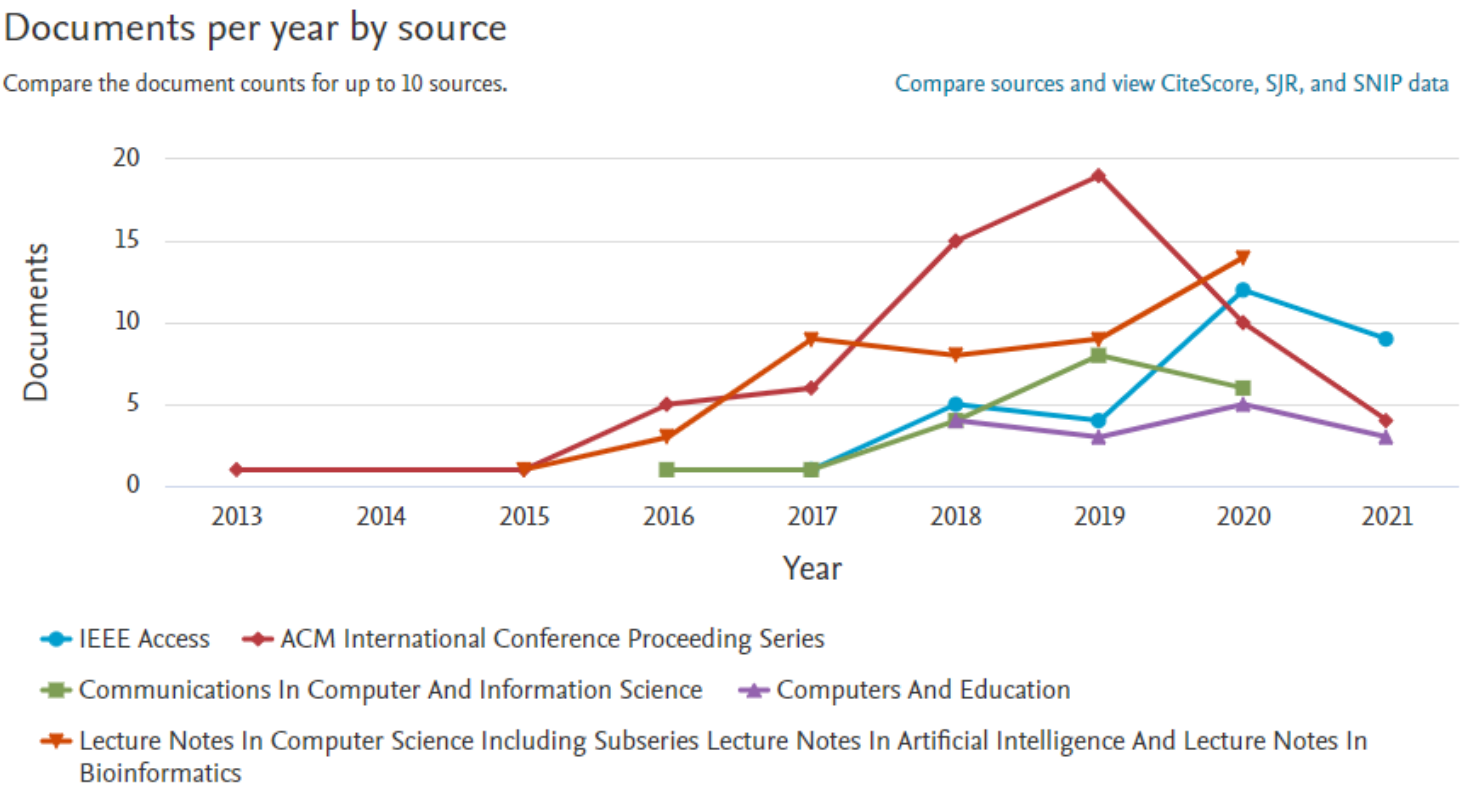

Figure 3

Attributes researched the most over the years 2013-2021

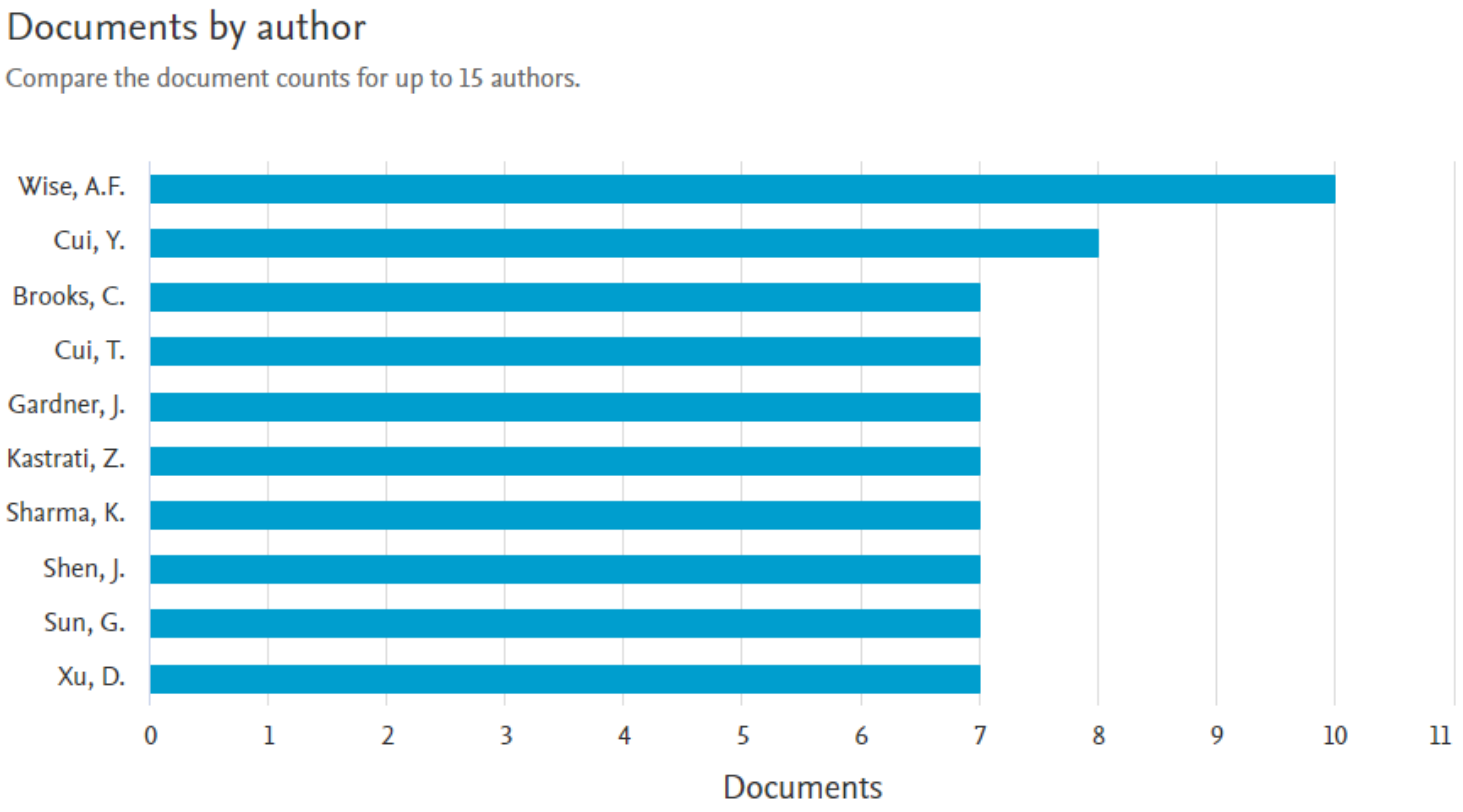

Figure 4

Comparative review of documents by authors

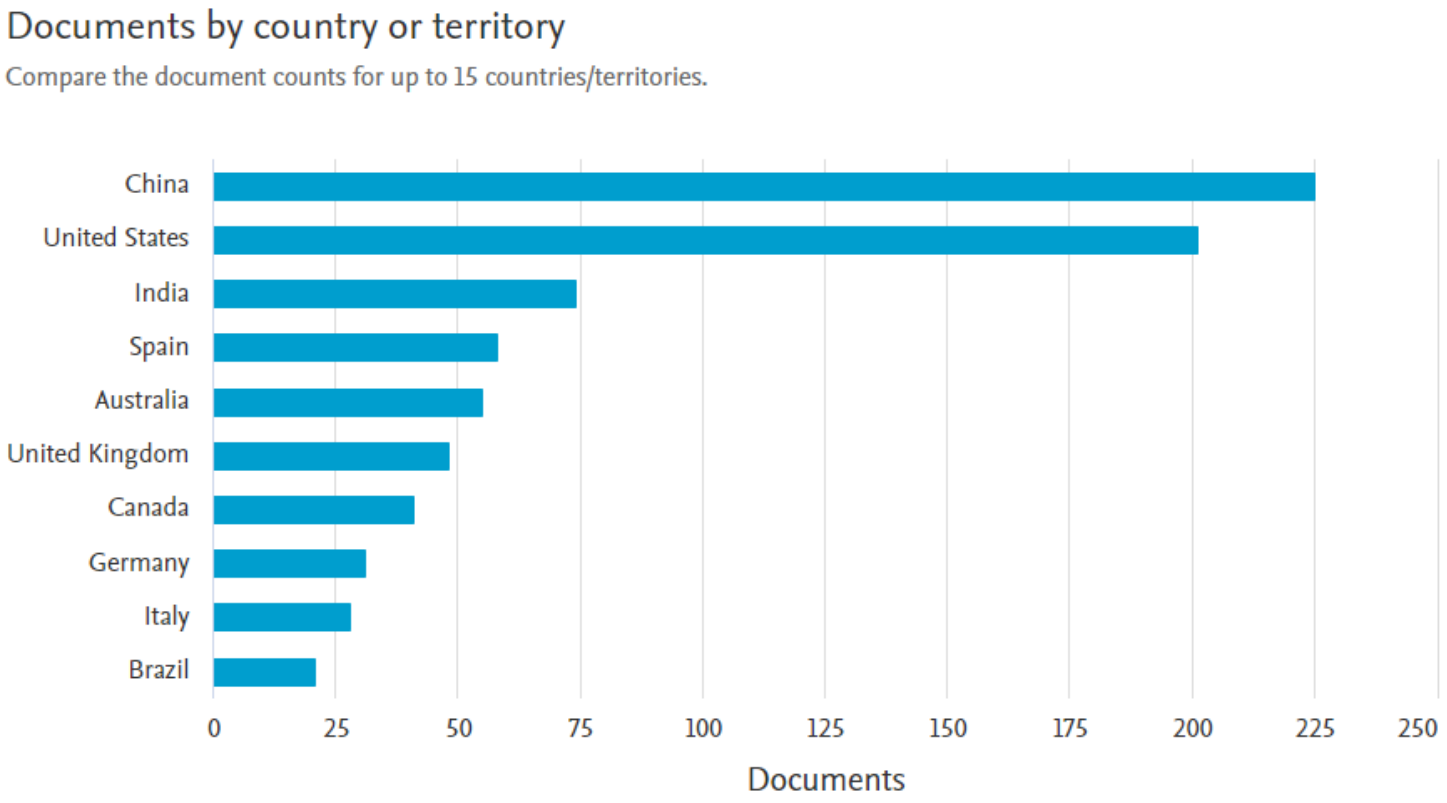

Figure 5

Publication of documents by country

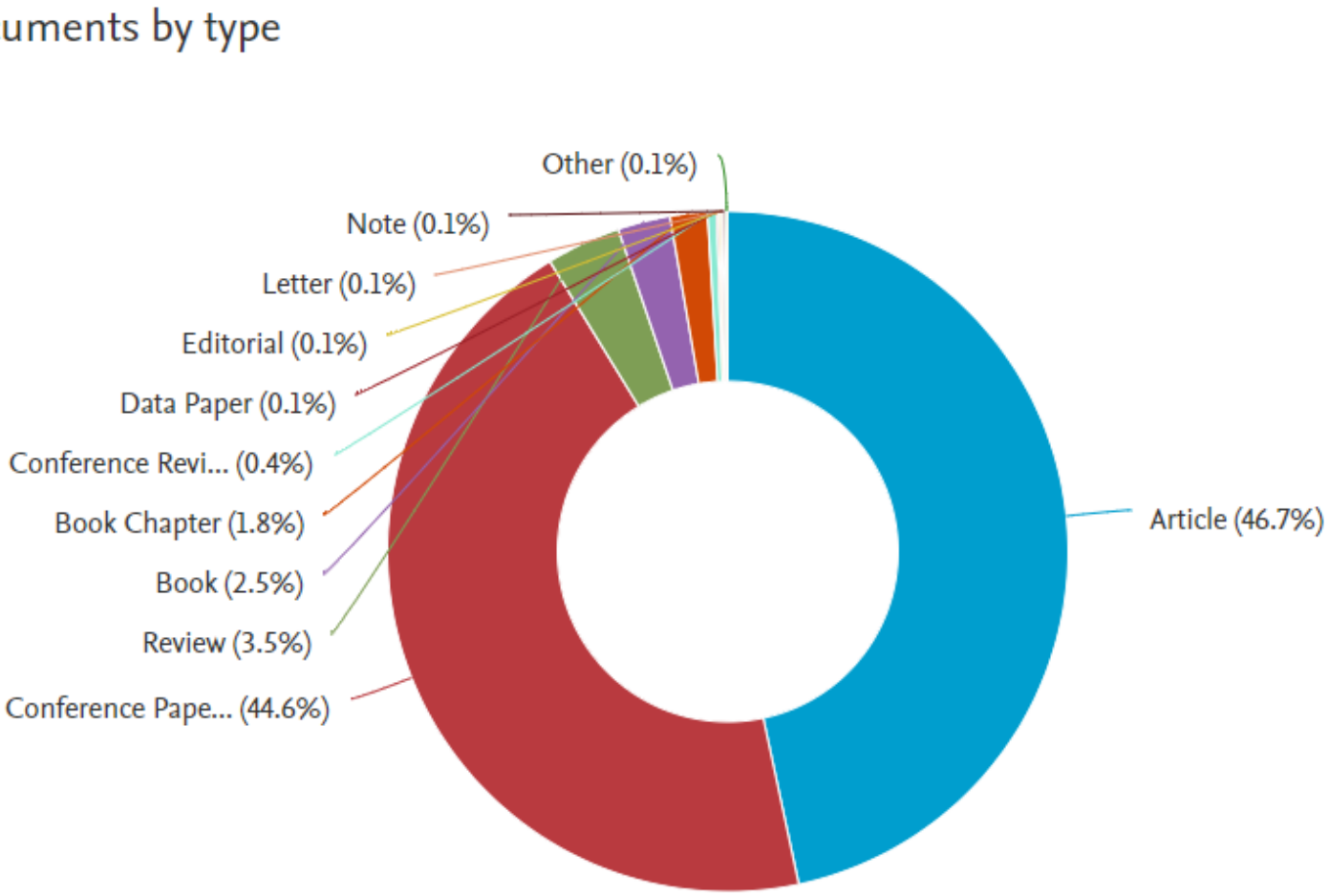

Figure 6

Publication of documents type

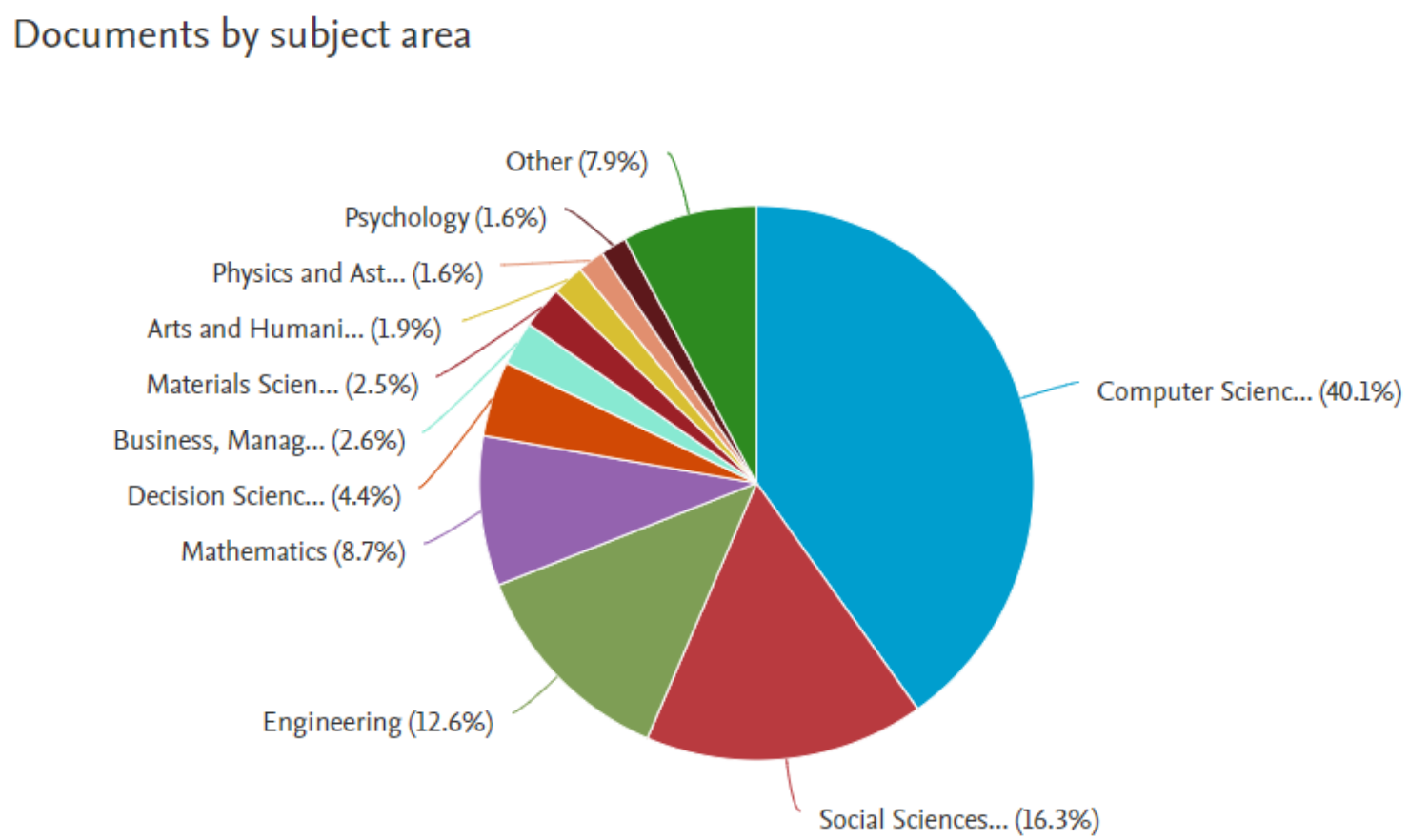

Figure 7

Subject wise published document

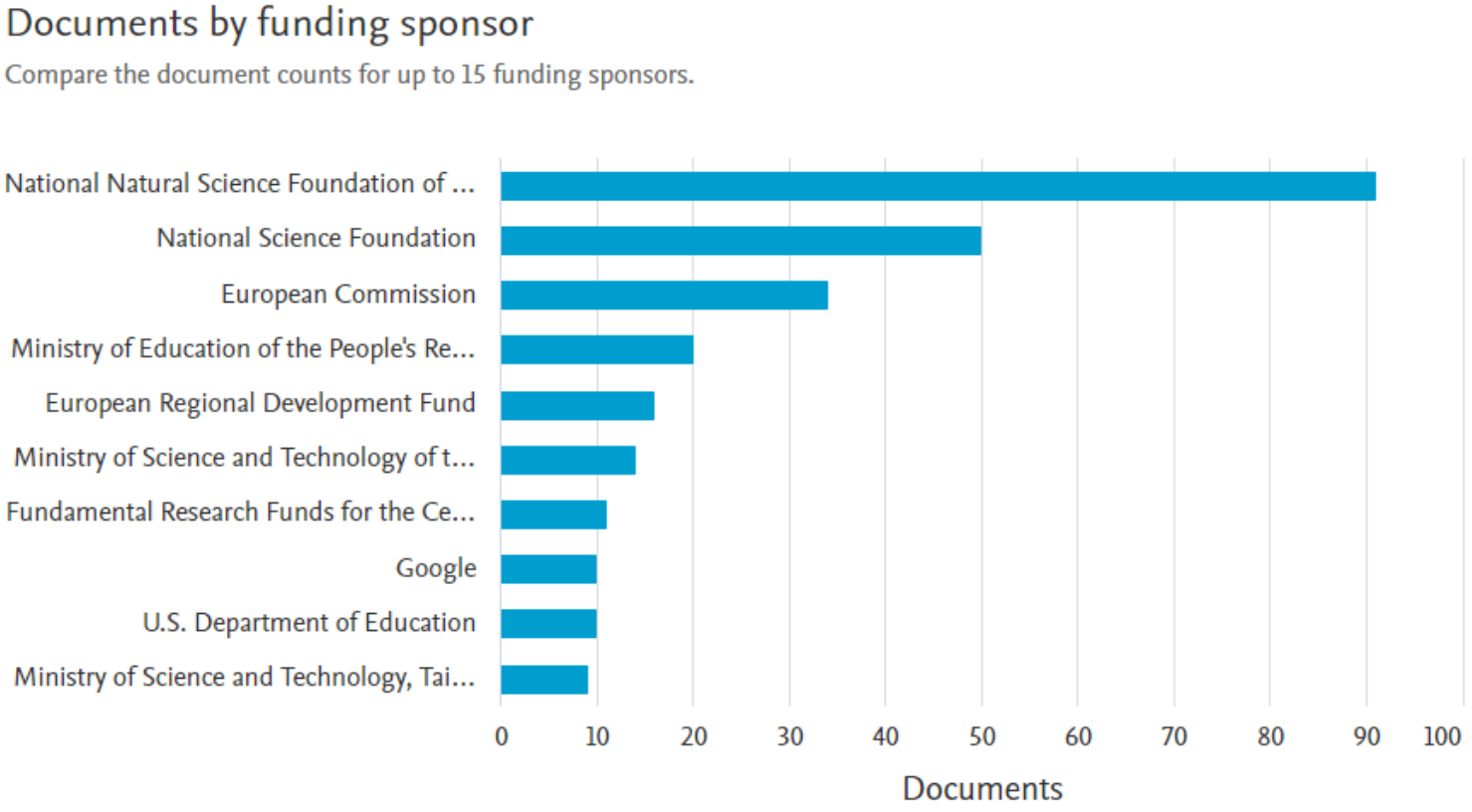

Figure 8

Funding's for the research work

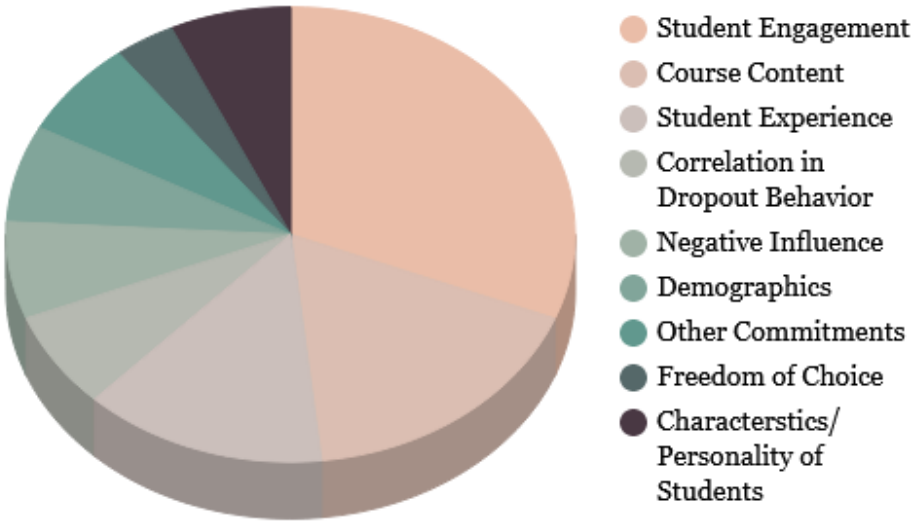

Figure 9

Attributes researched the most over the years 2013-2017