MOOC Dropouts: A Multi-System Classifier

Massimo Vitiello[1], Simon Walk[2], Vanessa Chang[3], Rocael Hernandez Rizzardini[4], Denis Helic[1], and Christian Guetl[1]

[1] Graz University of Technology massimo.vitiello@student.tugraz.at, dhelic@tugraz.at, cguetl@iicm.edu
[2] Stanford University walk@stanford.edu
[3] Curtin University of Technology vanessa.chang@curtin.edu.au
[4] Universidad Galileo roc@galileo.edu

Abstract. In recent years, technology enhanced learning platforms became widely accessible. In particular, the number of Massive Open Online Courses (MOOCs) has—and still is—constantly growing. This widespread adoption of MOOCs triggered the development of specialized solutions, that emphasize or enhance various aspects of traditional MOOCs. Despite this significant diversity in approaches to implementing MOOCs, many of the solutions share a plethora of common problems. For example, high dropout rate is an on-going problem that still needs to be tackled in the majority of MOOCs. In this paper, we set out to analyze dropout problem for a number of different systems with the goal of contributing to a better understanding of rules that govern how MOOCs in general and dropouts in particular evolve. To that end, we report on and analyze MOOCs from Universidad Galileo and Curtin University. First, we analyze the MOOCs of each system independently and then build a model and predict dropouts across the two systems. Finally, we identify and discuss features that best predict if users will drop out or continue and complete a MOOC using Boosted Decision Trees. The main contribution of this paper is a unified model, which allows for an early prediction of at-risk or dropout users across different systems. Furthermore, we also identify and discuss the most indicative features of our model. Our results indicate that users' behaviors during the initial phase of MOOCs relate to their final results.

1 Dropouts and At-Risk Users in MOOCs

With a widespread access to the Internet, education has evolved remarkably. Massive Online Open Courses (MOOCs), which can potentially reach audience at a global scale emerged as an option to acquire knowledge, as they also exhibit significant advantages for both users and content creators. The majority of MOOCs on the Web are freely available and have no entry requirements, which further encourage enrollments [17] [21]. Over time, platforms such as edX[5], Coursera[6] and Udacity[7], developed a monetisation model around this emerging

[5] https://www.edx.org/
[6] https://www.coursera.org/
[7] https://www.udacity.com/

ecosystem. The idea of obtaining a certificate after completing a MOOC in exchange for a small fee has already proven to be an appealing option for users, acknowledging their time, efforts and achievement.

Issue. Despite their massive appeal, MOOCs are known to suffer from high dropout rates. This is a particularly pressing issue, as on average about 90% of all enrolled users do not complete their classes [14]. Early detection of at-risk users, who are nevertheless eager to successfully complete a course, is very important. This would allow operators of MOOCs to devise strategies to intervene and mitigate the number of *dropouts*, those at-risk users who eventually abandon a course. Moreover, studies on MOOCs dropouts generally focus on very specific domains, with well-structured courses, characterized by assignment deadlines and fixed course lengths. What has been missing up to now is a study or a baseline that compares factors that influence the dropout rates across different MOOC systems and layout of MOOCs.

Motivation. Hence, it is important to identify features that are best suited to predict potential dropouts at an early stage. This would give MOOCs' providers actionable information, allowing them to adapt their courses accordingly. Additionally, comparing features that best distinguish completers and dropouts across different systems will yield new insights into general behavioral patterns that dictate individual outcomes of MOOCs for online learners. Specifically, the identification of such features, common to MOOCs across different systems, can reveal useful information to devise new strategies to mitigate the high dropout rates and keep users engaged and motivated when participating in MOOCs.

Approach. First, we conduct and evaluate prediction experiments to detect at-risk users in early stages of MOOCs from two different systems. Second, we train a model based on features present in all our datasets, to identify the best predictors of dropouts across different systems. Third, we conduct all of our experiments with a varying number of interactions, allowing us to measure if the ranking and importance of the features change over time. Finally, we discuss the implications of our findings in the context of the different experiments.

2 Related Work

Analyzing MOOCs and Features. Traditionally, analyses involving MOOCs are carried out by first identifying groups of users based on the similarity of their expectations and goals at the point of enrollment. A foundation for all of these studies is the *Funnel of Participation* [5]. In this study, the process towards completion of a course is composed of 4 phases: awareness, registration, activity and progress. Each of these phases is characterized by a certain *attrition* of the number of active users. Further studies analyzed users' surveys to understand reasons for drop out, detailing the attrition as either *healthy* or *unhealthy* [8] [11]. Server logs were also analyzed for users classification by means of clustering approach [15] [16] and linear regression model [6]. Features' importance and their mutual interactions, were studied in relation to machine learning algorithms, such as Support Vector Machines (SVM) [3] [4] and Decision Trees [10].

Detecting Dropouts. Many researchers dealt with dropout classification by means of log analysis and machine learning. Jiang et al. [13] applied a logistic regression model on a four weeks MOOC offered on Coursera. They tried to predict if users would obtain a certificate and if it would be a normal or a distinction one. Their findings indicated that the first-week assignment scores were a strong indicator of users' performance at the end of the course. In Xing et al. [20] the authors proposed a model to predict whether a user will drop out in the following week. Their results indicated that weekly features were more effective than the cumulative ones. Boyer and Veeramachaneni [2] experimented with dropout prediction in a real-time scenario. They used a rolling window, whose size represented the number of past weeks which they considered to construct features. Their results suggested that using a lower amount of past information could yield results comparable to the ones from a full window size.

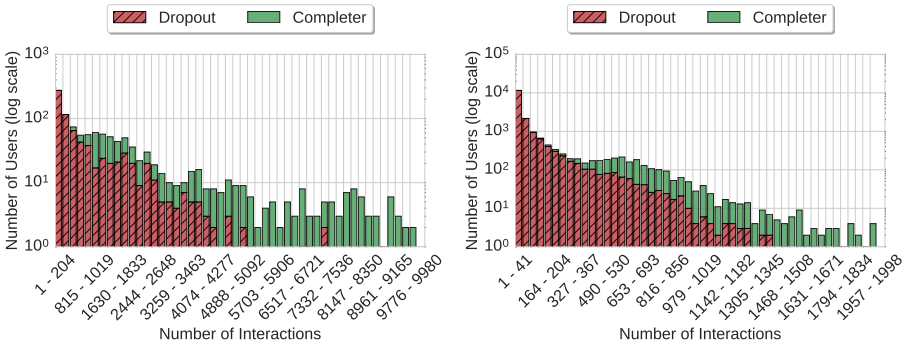
Balakrishnan and Coetzee [1] used Hidden Markov Models (HMM) to predict if users will drop out in the following week. The dataset consisted of a MOOC from Berkeley University, offered on edX. Their results can be used to suggest changes in the engagement style to those students who are more likely to drop out in the close future. Vitiello et al. [19] attempted dropout predictions over a set of 5 MOOCs. Their results indicated that certain combinations of features could significantly improve prediction scores. In Sinharay [18], the author presented a detailed review of different data mining techniques and compared their performance with real-data examples. Particularly, the author predicted dropouts on a dataset including students from various high schools in Florida. The obtained results indicated that methods such as Random Forests and Boosting can improve performance in regard to linear and logistic regression approaches.

The work presented in this paper further extends the state-of-the-art by analyzing MOOCs from two different sources: Universidad Galileo and Curtin University. We initially analyze and perform a dropout prediction experiment on each of these individually. Then, we excerpt a multi-systems model for MOOC evaluation and classification of users likely to drop out. In order to do so, we rank our features according to their importance and compare the obtained results.

3 Materials and Methods

3.1 Dataset

Table 1 shows the characteristics of the MOOCs from Universidad Galileo and Curtin University. Logs of Curtin University include interactions of each enrolled person, while those of Universidad Galileo only report interactions of active users (or learners). In our setting, interactions coincide with clicks of users in the MOOCs' environment.In particular, a total of $3,157$ active users in our datasets are from Universidad Galileo, and $35,473$ enrolled users are from Curtin University. We will use the more general term *users* to refer to these groups for each system. The average number of interactions for MOOCs of Curtin University is significantly lower than the ones of Universidad Galileo (see Figure 1). We can



(a) Universidad Galileo (b) Curtin University

Fig. 1. Number of interactions for each class (dropout vs. completer). Figure 1(a) refers to Universidad Galileo and Figure 1(b) to Curtin University. The number of interactions are grouped in bins and reported on the x-axis, while the y-axis represents the amount of users (in log scale). Completers are plotted in green and Dropouts in red. For both systems, Dropouts are present in higher number and are less active than the Completers.

see that Completers interact more often with the MOOCs in both systems, while the percentage of Dropouts is higher for the datasets from Curtin University.

Table 1. Characteristics of all MOOCs. The analyzed MOOCs belong to 2 different systems, Universidad Galileo and Curtin University. MOOCs of Universidad Galileo have a fixed schedule and are characterized by lower number of active users and restrained dropout rates. In contrast, MOOCs of Curtin University are self-paced and account for a higher number of enrolled users and dropout rates. The average number of interactions of Curtin University's MOOCs is lower than the ones from Universidad Galileo's MOOCs.

System	MOOC Title	Users	Completers	Dropouts	Dropout Rate	Average Interactions Global	Completers	Dropouts
Universidad Galileo	Android (AND)	583	77	506	87%	433	1597	260
	Authoring tools for E-Learning (AEL)	255	101	154	60%	722	1401	279
	Client Attention (CA)	89	60	29	33%	394	510	154
	Cloud Based Learning (CBL)	274	121	153	56%	2353	4423	747
	Community Manager (CM)	811	320	491	60%	850	1760	268
	Digital Interactive TV (DITV)	117	63	54	46%	999	1582	319
	Introduction to E-learning (EL)	239	81	158	66%	1623	3804	545
	Medical Emergencies (ME)	118	49	69	59%	1671	3172	606
	User Experience (UE)	182	62	120	66%	499	1137	170
	Web Tools and Educational Applications (WTEA)	176	99	77	44%	265	369	131
	Web Tools in the Classroom (WTC)	313	131	182	58%	1044	2078	299
Curtin University	MOOCC1	21948	1500	20448	93%	93	683	49
	MOOCC2	10368	208	10160	98%	58	760	44

MOOC Systems. It is important to understand that the MOOCs are hosted and implemented on two different systems, and, therefore, the structure and organization radically differ. In particular, Universidad Galileo's courses are organized with a predetermined schedule and calendar, where each MOOC lasts

8 weeks, including assignments. In contrast, the courses from Curtin University are organized in a self-paced mode; after a MOOC's official start, all the materials would be available online to the enrolled users, who would then participate and engage at their own pace. Moreover, there are no deadlines for the assignments and the duration of the MOOCs is generally flexible. Universidad Galileo's MOOCs are implemented for experts with a technical background in a particular field, who want to further develop their knowledge. The ones from Curtin University, however, are intended for a more general audience, not necessarily with experience on the topic of the course.

Feature comparison. Additional differences include the tools that each system deployed/implemented and the level of granularity of the logged interactions for later analysis. Aside from common information, such as *Timestamp* and *User id*, interactions in Universidad Galileo's logs would fall into one of 20 categories: *Assessment*, *Assignment*, *Evaluation*, *File Storage*, *Forum*, *Learning Content*, *Peer Evaluation*, *Calendar*, *Course Members* among others. On the other hand, the MOOCs offered on edX by Curtin University provide more detailed logs of interactions[8]. In particular, requests are divided into 7 different macro-groups (as shown in Table 2). EdX logs from Curtin University also include *Enrollment* interactions, which indicate enrollments for both users and course instructors. We use these interactions to identify the total amount of enrolled users, but we do not consider such interactions when constructing the features.

Multisystem dataset. We create three additional datasets; the first one combines users of all MOOCs of Universidad Galileo, the second one includes users of all MOOCs of Curtin University and the third one combines users of all MOOCs from both systems. We reference them as *Galileo*, *Curtin* and *MIX* respectively. We use these datasets to predict dropouts on a system-to-system and multisystem level. Moreover, we use these to analyze the importance of the features.

Feature extraction. A *feature* is a characterization of users' engagement in a MOOC that we regard indicative and helpful to identify dropouts. We describe each user in terms of a set of features, which is input to the classifier and summarize these in Table 2. These features can be split into two groups. The features within the first group consist of time-related information, obtainable for both systems. These features build up the concept of user' sessions, which are defined as a set of actions, where the timespan between each action is less or equal than 30 minutes. The second group of features is system dependent.

3.2 Prediction Model

We first outline the steps for the proposed experiments and then describe each of these in detail.

Feature Extraction. The sooner we can predict if a user is likely to drop out, the earlier we can develop strategies to intervene and engage with the user. Hence, we focus on users' initial interactions and construct the features described in Section 3.1, following two different strategies. First, we focus on the initial

[8] A complete description of edX logs can be found at http://edx.readthedocs.io

per-user absolute interactions. We set up our experiments ranging from 1 to 100 per-user absolute initial interactions, on which we calculate the features. Secondly, we consider the number of interactions taking place in the first week after users' first interaction with the MOOC. In this case, we determine the timestamp of a users' first interaction and consider all interactions that take place within a certain timespan (1 to 7 days).

Class Balancing. For both systems, the number of Dropouts is significantly higher than the number of Completers. We addressed this class imbalance problem by oversampling [9] [12]. This means that new samples are randomly picked and added to the class with fewer examples until its dimension equals to the one of the larger class.

Training. Once classes are balanced, we split the examples into a training set, used to train the classifier, and a test set, which we use for evaluation. We use a ratio of 80:20 between training and test datasets, using a Stratified Shuffle Split with 10 folds. With this approach, each fold will also be balanced in the number of examples from each class. Furthermore, the shuffle assures that each fold will consist of different examples.

Evaluation. Finally, we use accuracy to evaluate the prediction error of the experiments. Accuracy is defined as the fraction of correctly predicted examples and is therefore bounded between 0 and 1. An accuracy of 0 means that every example has been misclassified, while an accuracy of 1 indicates that every example has been correctly classified. Furthermore, we run the experiments for each fold until the mean prediction error converges.

3.3 Dropout Classification

We are interested in understanding the reasons that lead users to drop out at a certain point and to assess the number of interactions that are necessary

Table 2. Feature Description. We consider 3 kind of features, one is common to both systems and the other two are system dependent. *Temporal* features are derived from users' sessions and are used with both systems. *Tool* from Universidad Galileo includes 20 different features that map to the tools available for this system. *Tool* from Curtin University accounts for 7 different groups (MOOCs' components), each of these comprising a wide range of interactions, for a total of around 100. For both systems, we calculate these features counting the number of interactions that belong to each tool.

Type	Feature	Domain	Description
Temporal	Sessions & Requests	Both	Total number of sessions and of requests
	Active Time & Days	Both	Total amount of active time and of active days
	Timespan Clicks	Both	Average timespan between two consecutive clicks (within same session)
	Session Length & Session Requests	Both	Average session lengthh and requests per session
	Active Days Requests	Both	Total number of requests for each active day
Tool	Requests per Tool	Universidad Galileo	Total requests per each tool (ex. *Evaluation*, *Assignment*, *Forum*)
Tool	Course Navigation	Curtin University	Interactions within the course content page (ex. *Link Clicked*, *Tab Selected*)
	Video	Curtin University	Interactions with video components (eg. *Play Video*, *Show/Hide Transcript*)
	Problem	Curtin University	Interactions with the problem module (eg. *Problem Grade*, *Show Hint*)
	Poll & Survey	Curtin University	Interactions with the Poll and Survey block (eg. *Submit*, *Show Results*)
	Bookmark	Curtin University	Interactions with the Bookmark component (eg. *Add/Remove Bookmark*)
	Discussion Forum	Curtin University	Interactions happening within the Forum (eg. *Search*, *Comment*, *Vote*)
	Main Page Links	Curtin University	Clicks on main page links (ex. *Progress*, *Instructor*, *Study at Curtin*)

to identify potential dropouts and if different features yield equivalent results. Furthermore, we want to compare different MOOCs and systems to check for similarities and differences.

Initially we run prediction experiments on each MOOC independently (see Figure 2) before comparing the results between systems (see Figures 2(c) and 2(d)). For the individual prediction experiments we use Support Vector Machines (SVM), which try to find the optimal hyperplane (in higher dimension spaces) to separate data points.

For the system-to-system and multisystem experiments, we predict dropouts using Boosted Decision Trees [7]. This ensemble classifier combines the outputs from a set of single decision tree in a sequential way. For each learned model, the examples are re-weighted; the misclassified ones receive a higher weight, while the correctly classified ones get a lower weight. This way, the next decision tree will focus more on the misclassified examples. Overall, we propose three experiments. First, we conduct two system-to-system dropout prediction experiments. We use the *Curtin* dataset for training and the *Galileo* dataset to test our classifier. Second, we switch the datasets and train on *Galileo* to predict dropouts on *Curtin*. We denote these experiments as *Curtin on Galileo* and *Galileo on Curtin* respectively. Third, we use the *MIX* dataset, in which the training and test sets include examples from both systems. Third, we use the *MIX* dataset, in which the training and test sets include examples from both systems. Finally, we determine the importance scores for our features from the Boosted Decision Trees to identify the predictive power of each feature for the detection of dropouts.

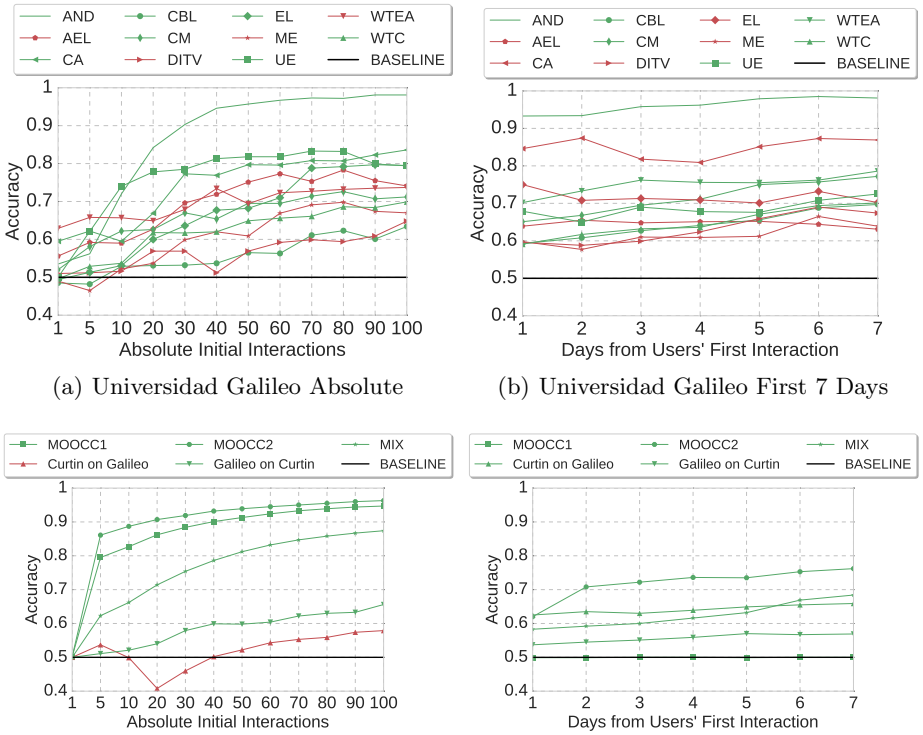
4 Results

4.1 Dropout Classification

Figure 2 depicts the mean (over the 10 folds) accuracy for each MOOC. The y-axes reports the accuracy and the x-axes indicates the number of absolute interactions (Figures 2(c) and 2(c)) and the considered number of days from the users' first interaction (Figures 2(d) and 2(d)).

Universidad Galileo. As shown in Figure 2(a) and 2(b), for MOOCs of Universidad Galileo, we see that not always increasing the number of considered interactions and days guarantee higher accuracy. Firstly, there is a set of MOOCs plotted in green, for which the accuracy increases over time or, after an initial growth, stabilizes. The second group is plotted in red and consists of MOOCs for which the accuracy trend is less steady. For the Absolute Experiment, except for the *AND* MOOC, the first 100 users' absolute interactions are too few for a correct classification of the users. For the First 7 Days Experiment, the increase in accuracy is less significant in respect of the Absolute Experiments. In some cases, as for the *CA* and *AEL* MOOCs, considering more days can lead to a worsening of the accuracy. With the exception of *AND*, these two approaches do not guarantee a precise detection of dropouts over the MOOCs in this system.

Curtin University. Figure 2(c) and 2(d) report the results for Curtin University's MOOCs. For the Absolute Experiment, we obtain for both MOOCs an



(a) Universidad Galileo Absolute

(b) Universidad Galileo First 7 Days

(c) Curtin & Multi-System Absolute

(d) Curtin University & Multi-System First 7 Days

Fig. 2. Single SVM and Multi-System Boosted Decision Tree Results. Figures 2(a) and 2(b) report the accuracy results for Universidad Galileo in relation to the absolute number of interactions and the first 7 days from users' first interaction metrics (for MOOCs label explanation see Table 1). The results for these metrics for Curtin University and the multi-system experiments are depicted in Figure 2(c) and 2(d) respectively. Accuracy of the MOOCs plotted in green is increasing or becoming stable after a certain point. MOOCs plotted in red are not characterized by such trend.

accuracy always higher than 0.8 already with only 5 absolute interactions. We investigate further these situations and find out that the most used tools with 5 interactions belong to *Video* and *Main Page Links* components. The higher the amount of absolute considered interactions is, the more the users engage with *Video*, *Course Navigation* and *Problem*. *Discussion Forum* is only rarely used, mostly for visualization purposes. Therefore, we conclude that there is a particular set of components that strongly catalyze users' attention. Results for the First 7 Days Experiment have a generally low accuracy. The accuracy for *MOOCC2* has a slightly increase the more days are considered, while for *MOOCC1* the accuracy is steady at 0.5. These low accuracy values can be due

to the Course Enrollment interactions (see Section 3.1) that introduce a certain noise in this setting.

Multi-System. Figure 2(c) and 2(d) also report the accuracy for the three multi-system experiments. For the Absolute Experiment, the accuracy of *Curtin on Galileo* increases steadily when the interactions considered are more than 20. The accuracy for *Galileo on Curtin* and *MIX* instead, is always increasing. Particularly, the accuracy for *MIX* resembles the ones of *MOOCC1* and *MOOCC2*. For the First 7 Days Experiment, we notice a slight increase in the accuracy the more days are considered for all the experiments. The accuracy profile for *MIX*, is the one with the broader increase the more days are taken into consideration, while *Galileo on Curtin* is the setting that yields the higher accuracy. For *Curtin on Galileo* the accuracy remains almost unaltered. Generally, we obtain better results with the absolute number of interactions approach. The difference in the accuracy is particularly marked for the *MIX* dataset.

Findings. For self-paced MOOCs, such as those from Curtin University, a small number of initial interactions contains already valuable information for a correct classification of the users. Particularly, given the high details of the logs, the number of interactions with each tool is a strong indicator whether users will drop out or not. Due to users' enrollment actions, the first 7 days from users initial interaction reveals to be a less effective approach for self-paced MOOCs. For fixed schedule MOOCs, such as those from Universidad Galileo, both metrics are less accurate. This may be partly due to the structure of the courses.

4.2 Features Analyses

Considering the results for the absolute interactions experiment, we can split the features into 2 groups; a group of high scoring features, consisting of *Session Length*, *Timespan Clicks*, *Requests* and *Active Time*, and a group of low scoring ones, including *Days*, *Active Days Requests* and *Sessions*. We note that, for the high-scoring group, there is no feature that always outperforms the others. Moreover, this group division is present across all 3 experiments despite the considered number of interactions. We conclude that using the initial 100 users' interactions as a metric, we can clearly identify the features that best split the users between Completers and Dropouts. Also for the first 7 days from users' first interaction experiment, we can still split the features in high and low scoring ones. For *Curtin on Galileo* and *MIX* experiments, the high scoring features group consists of *Timespan Clicks*, *Requests* and *Active Time*. These remain unmodified in respect to the considered days. For *Galileo on Curtin* the scoring seems to be less definite, with only *Session Length* always belonging to the high scoring ones. From this metric, we are able to identify a set of most valuable features.

Findings. Among the different multi-system experiments and the considered metrics, we identify two classes of features; high-scoring and low-scoring. Beside small variations, features always belong to only one of these classes. This implies that, despite the differences between the two systems, when they are analyzed

together, there are strong similarities regarding the importance of the features. Moreover, *Days*, *Sessions* and *Active Days Requests* are always the features with lowest weights. The remaining features represent a set with high weights for both metrics and across the systems.

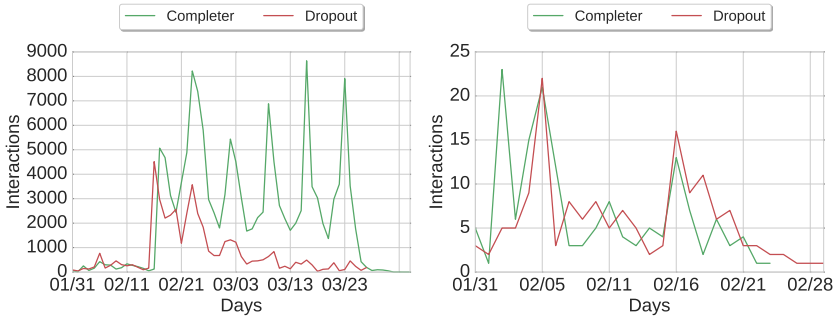
5 Discussion

5.1 Dropout Classification

Curtin University's MOOCs are characterized by a steady increase in accuracy the more interactions or days are considered and by an accuracy higher than 0.8 for 5 absolute interactions. Such an increase in accuracy is not always present for MOOCs from Universidad Galileo. Except for *AND*, which has an accuracy profile similar to the MOOCs from Curtin University, MOOCs from Universidad Galileo rarely have an accuracy of 0.8 or higher. Reasons of this discrepancy in the accuracy could be due to differences in the didactic settings, including the structure of the course and type of activities between the systems. First, Universidad Galileo's MOOCs, although having a defined 8 week duration, are sometimes subjected to a later start. This happens, for example, when a MOOC is accessible to the users but the material is not yet available on the platform. In this situation, there is an initial phase characterized by few interactions (see Figure 3(a)), followed by a burst of activity of Completers and Dropouts, once either the material has become available or the MOOC officially started (see Figure 3(b)). The lower accuracy values for some of the MOOCs from this system, can be a consequence of these particular situations.

On the other hand, Curtin University's MOOCs are organized in a self-paced manner, with the entire material and resources available to users from the start. Furthermore, *MOOCC1* and *MOOCC2* have an average number of interactions of 93 and 58 respectively (see Table 1). This means that the first 5 interactions of each user represent, on average, 5.38% and 8.62% of their total interactions for *MOOCC1* and *MOOCC2* respectively. These percentages are much higher than those from Universidad Galileo's MOOCs; the highest for this system comes from *WTEA*, for which 5 interactions represent on average only 1.89% of a users' total interactions. Therefore, the considered number of absolute interactions is too low for Universidad Galileo.

Similarly, these situations can be also observed when considering the first 7 days after a users first interaction. As previously mentioned, the lack of interactions in the initial phase of Universidad Galileo's MOOCs, could be due to delays with uploading of materials and the official start. In the first case, it is possible that users do not interact with the MOOC in the successive days. It is more likely that only when a MOOC's material becomes available, users will again engage with the MOOC. Thus, it is possible that considering only the first 7 days from a users first interaction will only add a few extra interactions. The results for the MOOCs from Curtin University are presented in Figure 2(d). These are generally worse than those from the absolute experiment, particularly for *MOOCC1*, where the accuracy is constant at 0.5. We believe that users who only sign up for



(a) Interactions per Class (b) Users' First Interaction per Class

Fig. 3. MOOC AEL Interaction per Class. Figure 3(a) reports distribution of all interactions per class. Interactions antecedent 02/16 took place during a phase where probably course's material was not available or the MOOC did not start yet. The distribution for Completers and Dropouts is similar during this time. Figure 3(b) shows users' first interaction for each class. Most first interaction happened before 02/16, date in which there is an increase for both classes. Some Dropouts firstly interacts with the MOOC more than one week after the 02/16.

a MOOC, but never interact with it, or potentially interact with it at very late stages of the course could potentially influence the prediction. From Curtin University's logs we can extract a total of $8,552$ Dropouts with only one interaction for the MOOC *MOOCC1*, and a total of $4,436$ for *MOOCC2*. Furthermore, if we consider only the active users, by completely dropping the Course Enrollment actions together with these Dropouts and re-run the experiments we obtain the results as shown in Figure 5.1. These results are much more in line with those obtained for the absolute number of interaction experiments. For *MOOCC1* and *MOOCC2*, users' first day of interactions, is sufficient to achieve an accuracy of 0.9, which steadily increases when more days are considered.

The multi-system experiments, using Boosted Decision Trees, also benefit of the removal of Course Enrollment actions. *Galileo on Curtin* and *Curtin on Galileo* have values for the accuracy higher than the ones in the absolute interaction experiments. For *Curtin on Galileo* the accuracy increases when more days are considered, while for *Galileo on Curtin* it lowers slightly for 6 and 7 days. This may be caused by an initial phase with a low number of interactions in Universidad Galileo's MOOCs (see Figure 3(a)), which introduces noise for the classifier. However, the accuracy increases for the prediction experiment using the *MIX* dataset. Already the first day of interactions is sufficient for an accuracy of 0.7.

5.2 Feature Analyses

For the Absolute Experiment the group of high scoring features includes *Session Length*, *Timespan Clicks*, *Requests* and *Active Time*. From these, the weights

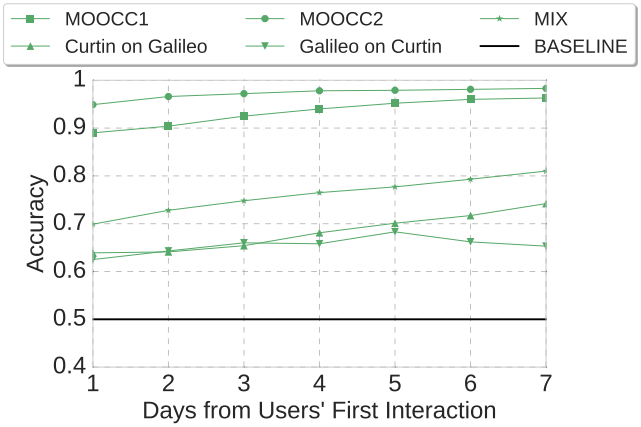


Fig. 4. Single SVM and Multi-System Boosted Decision Tree Results without Enrollments. Discarding the enrollment actions, yields better results. This is due to Curtin University's MOOCs schedule, for which an enrollment phase of up to a couple of months precede the official start of the MOOCs.

for *Requests* are the highest when 100 absolute interactions are considered. This is reasonable for MOOCs with predefined schedule as those from Universidad Galileo, in which users are forced to keep up with a certain pace according to deadlines, exams and assignments. The high score of this feature for all multi-system experiments, seems to imply that this is true also for self-paced MOOCs from Curtin University. Although *Sessions* is one of the less valuable features, the (average) session characteristics, such as *Session Length*, *Timespan Clicks* and *Active Time*, have higher scores. This suggests that users' behavior during a session relates stronger to whether users are Completers or Dropouts, than the number of sessions they have. For the First 7 Days Experiment, the rankings of *Curtin on Galileo* and *MIX* are similar to those obtained in the Absolute Experiment, with *Timespan Clicks*, *Requests* and *Active Time* always being the features with highest weights. For *Galileo on Curtin*, *Session Length* and *Requests* are almost always the highest scoring features. This mixed ranking could be due to the smaller dimension of Universidad Galileo's dataset, in respect to Curtin University's one. However, this aspect does not seem to be relevant for the Absolute Experiment. We can conclude that, features constructed considering up to the first 7 days after users' first interaction, do not relate to users dropping out, as much as those obtained from users' initial absolute interaction. This claim is supported by the results of Figure 2, where for Universidad Galileo's MOOCs the increase in accuracy for the First 7 Days Experiment, is more moderate than the one from the Absolute Experiment. The features *Active Days*, *Sessions* and (with one exception) *Requests Active Days* are always the lowest scoring for all experiments in the multi-system scenario.

6 Conclusion

With this work, we faced the problem of early classification of at-risk users in MOOCs. To address this shared problem, we analyzed MOOCs from two different systems in a homogeneous way, using Support Vector Machine and Boosted Decision Tree. We investigated two aspects, the initial absolute number of users interaction and the first 7 days after users' first interaction with the system. We obtained the best results when up to the first 100 absolute interactions were considered. For Curtin University's MOOCs we identified a set of components mostly used by the users, that strongly indicates whether users will drop out or not. Particularly, we verified that interactions with *Video* and *Course Navigation* components are representative of user engagement even during the very initial phase of the course. We also discovered that other components (*Discussion Forum* primarily) are only marginally important and scarcely used. Furthermore, we proposed a model for early dropouts detection in a multi-system setting. Despite the differences in the systems' structure (self-paced vs fixed schedule), topic, intended audience and conceptualization, we constructed a set of features shared by both systems.

In our future work, we will extend our model by enlarging the number of common features between the various systems. Further, we will conduct analyses with alternative approaches, which will also help to grasp and discover further aspects of the systems that we did not consider in this work. Moreover, we aim at further characterizing Completers and Dropouts by verifying if subgroups of users exist and experiment with users classification in a multi-class scenario.

7 Acknowledgments

The authors would like to thank the MOOC Maker Project[9], Universidad Galileo and Curtin University for providing the datasets for the analysis and the Graz University of Technology and Curtin University for supporting the research visits of Massimo Vitiello and Christian Gütl.

References

1. Balakrishnan, G., Coetzee, D.: Predicting student retention in massive open online courses using hidden markov models. Electrical Engineering and Computer Sciences University of California at Berkeley (2013)
2. Boyer, S., Veeramachaneni, K.: Transfer learning for predictive models in massive open online courses. In: International Conference on Artificial Intelligence in Education. pp. 54–63. Springer (2015)
3. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing multiple parameters for support vector machines. Machine learning 46(1-3), 131–159 (2002)
4. Chen, Y.W., Lin, C.J.: Combining svms with various feature selection strategies. In: Feature extraction, pp. 315–324. Springer (2006)

[9] http://www.moocmaker.org/

5. Clow, D.: Moocs and the funnel of participation. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. pp. 185–189. ACM (2013)
6. Coffrin, C., Corrin, L., de Barba, P., Kennedy, G.: Visualizing patterns of student engagement and performance in moocs. In: Proceedings of the fourth international conference on learning analytics and knowledge. pp. 83–92. ACM (2014)
7. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
8. Guetl, C., Chang, V., Hernández Rizzardini, R., Morales, M.: Must we be concerned with the massive drop-outs in mooc? an attrition analysis of open courses. In: Proceedings of the International Conference Interactive Collaborative Learning, ICL2014 (2014)
9. Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G.: On the class imbalance problem. In: Natural Computation, 2008. ICNC'08. Fourth International Conference on. vol. 4, pp. 192–201. IEEE (2008)
10. Guruler, H., Istanbullu, A., Karahasan, M.: A new student performance analysing system using knowledge discovery in higher educational databases. Computers & Education 55(1), 247–254 (2010)
11. Gütl, C., Rizzardini, R.H., Chang, V., Morales, M.: Attrition in mooc: Lessons learned from drop-out students. In: International Workshop on Learning Technology for Education in Cloud. pp. 37–48. Springer (2014)
12. Japkowicz, N., et al.: Learning from imbalanced data sets: a comparison of various strategies. In: AAAI workshop on learning from imbalanced data sets. vol. 68, pp. 10–15. Menlo Park, CA (2000)
13. Jiang, S., Williams, A., Schenke, K., Warschauer, M., O'dowd, D.: Predicting mooc performance with week 1 behavior. In: Educational Data Mining 2014 (2014)
14. Jordan, K.: Initial trends in enrolment and completion of massive open online courses. The International Review of Research in Open and Distributed Learning 15(1) (2014)
15. Kizilcec, R.F., Piech, C., Schneider, E.: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: Proceedings of the third international conference on learning analytics and knowledge. pp. 170–179. ACM (2013)
16. Li, N., Kidziński, Ł., Jermann, P., Dillenbourg, P.: Mooc video interaction patterns: What do they tell us? In: Design for Teaching and Learning in a Networked World, pp. 197–210. Springer (2015)
17. Liyanagunawardena, T.R., Adams, A.A., Williams, S.A.: Moocs: A systematic study of the published literature 2008-2012. The International Review of Research in Open and Distributed Learning 14(3), 202–227 (2013)
18. Sinharay, S.: An ncme instructional module on data mining methods for classification and regression. Educational Measurement: Issues and Practice 35(3), 38–54 (2016)
19. Vitiello, M., Walk, S., Hernández, R., Helic, D., Gütl, C.: Classifying students to improve mooc dropout rates. Research Track p. 501 (2016)
20. Xing, W., Chen, X., Stein, J., Marcinkowski, M.: Temporal predication of dropouts in moocs: Reaching the low hanging fruit through stacking generalization. Computers in Human Behavior 58, 119–129 (2016)
21. Yousef, A.M.F., Chatti, M.A., Wosnitza, M., Schroeder, U.: A cluster analysis of mooc stakeholder perspectives. RUSC. Universities and Knowledge Society Journal 12(1), 74–90 (2015)