Learning Analytics on Coursera Event Data: A Process Mining Approach

Patrick Mukala, Joos Buijs, Maikel Leemans, and Wil van der Aalst

Department of Mathematics and Computer Science
Eindhoven University of Technology, Eindhoven, The Netherlands
{m.p.mukala,j.c.a.m.buijs,m.leemans,w.m.p.v.d.aalst}@tue.nl

Abstract. Massive Open Online Courses (MOOCs) provide means to offer learning material in a highly scalable and flexible manner. Learning Analytics (LA) tools can be used to understand a MOOC's effectiveness and suggest appropriate intervention measures. A key dimension of such analysis is through profiling and understanding students' learning behavior. In this paper, we make use of process mining techniques in order to trace and analyze students' learning habits based on MOOC data. The objective of this endeavor is to provide insights regarding students and their learning behavior as it relates to their performance. Our analysis shows that successful students always watch videos in the recommended sequence and mostly watch in batch. The opposite is true for unsuccessful students. Moreover, we identified a positive correlation between viewing behavior and final grades supported by Pearson's, Kendall's and Spearman's correlation coefficients.

Keywords: Learning Analytics, MOOC, Coursera, Educational Data Mining, Process Mining, Online Learning

1 Introduction

Massive Open Online Courses (MOOCs) provide free learning opportunities to a wider online community. They are gaining considerable momentum and attract interests from accross different professions. This increasing interest in MOOCs triggers the need for continuous evaluation of their effectiveness, as many consider this learning approach to be in its infancy [8]. *Learning Analytics* (LA) [7, 11], attempts to provide useful insights pertaining to educational data [4].

Current LA literature indicates that the focus has been on classical data mining techniques to predict students' dropout risk and to estimate students retention rate etc., almost entirely using attributes such as students' age, previous grades, race, and academic qualifications etc. [4, 5]. This paper aims to analyze behavior. Therefore, classical statistical and data mining techniques are less appropriate. Instead, we use process mining [1].This adds a new perspective based on the actual behavior exhibited by students as they learn.

Process mining provides a set of algorithms, tools and techniques to analyze event data [1]. Three main perspectives offered by process mining include discovery, conformance checking and enhancement [1]. Discovery techniques allow the enactment of process models from log data. Conformance checking attempts to verify conformity to a predefined model and identify deviations, if any, while enhancement provides for models to be improved based on results of process discovery and conformance checking [1].

Given the online-based format and nature of MOOCs, it is possible to track students' activities following the individual clicks they make on the course webpages. The data generated this way can give us insights into how and when students follow lectures, and how they prepare for exams. Due to its popularity and availability of data, we analyze a MOOC hosted on the Coursera platform. Coursera keeps track of all students and staff activity details useful for our analysis. We extract and translate students' behavioral data into a sequence of events and analyze it in order to answer questions such as:

1. How do students watch lecture videos? What is the learning process?
2. What are the general viewing habits exhibited during this process?
3. What is the impact of such behavior/habits on the overall performance?

The remainder of this paper is structured as follows. We introduce our case study, the dataset used for analysis followed by a succint description of how events logs are derived from Coursera's data in Section 2. In Section 3, we describe students' learning behavior using process mining techniques: dotted chart and process discovery. In Section 4, we discuss learning behaviors following conformance checking alignments. In Section 5, we measure and discuss the correlation between students' learning behaviors and final grades. Section 6 concludes this paper and discusses possible future directions.

2 MOOC Data: The Coursera Case

Coursera subdivises raw data into three categories: general data, forums data and personal identification data. In total, the standard model comprises 59 tables storing information about users' privileges, announcements regarding the course, all forums details, assessements and evaluation data, course grades, submissions details etc. For the purpose of this study, we consider data obtained from Coursera for the first instance of the MOOC "Process mining: Data Science in action" which ran from November 11, 2014 to January 8, 2015. The overall statistics are detailed in Table 1.

We have limited our analysis to data about direct student behavior. The datasets we analyze are centered around the *students* participating in the MOOC, and the stream of *click events* they generated on the course webpages. A reference model for this selected part of the dataset can be found in [9].

Clickstream As students click on videos, they leave a trail of *click events*, called a *clickstream* associated with a particular lecture, or a particular quiz submission. In addition to the pages visited by a student (recorded as a *pageview ac-*

Table 1: Global statistics for our Coursera MOOC case study

Start date	**Nov 14, 2014**
# Registered	43,218
# Visited course page	29,209
# Watched a lecture	20,868
# Browsed forums	5,845
# submitted an exercise	5,798
# Certificates (normal/distinction)	1,688
# Normal certificate	1,034
# Distinction Certificate	654
End date	**Jan 8, 2015**

tion), we also know how the students interacted with the lecture videos (recorded as a *video action*).

Student For each student, we know the exact *time the student registered*, and if they participated in the special (paid) *signature track* or not. We also know if they get a *fail*, *normal* or *distinction grade.*

Course structure Lectures and quizzes are grouped into *sections* (*weeks*). Each section is visible to the students at a predetermined time (the *open time*). Within a section, lectures and quizzes may have their own open time, to guide students to follow a particular study rhythm. Finally, quizzes can also have deadlines (the *close time*).

Before we use process mining on this data for our analysis, we first need to build an event log. To do this, we need to decide which events are within scope and how events are grouped into cases. As an example, consider the sample data in Table 2. The resulting event log is shown in Table 3. Each student in Table 2 becomes one case in Table 3. For each case, we store the data available about the student, including their course grade data. For each clickstream event, we create an event belonging to the corresponding case (based on the student UserID). In this example, we only consider lecture pageview actions. For each clickstream event, we store the click event data, including the referenced lecture as event name.

3 Visualization of Learning Behavior

In this section we use the dotted chart and a process mining discovery algorithm (fuzzy miner [1]) to visualize the event data and discover the actual learning process. The aim is to visually provide insights on the overall MOOC and profile students' behavior throughout the duration of the MOOC. We consider three important dimensions in this analysis: the general lecture videos viewing habit, the quiz submission behavior as well as a combination of both. These insights can help to understand how students study and what impact such behaviors have on their involvement in the MOOC.

Table 2: Example of Student, ClickStream Event and Lecture data used to map cases and events in Table 3

Sample data about Student

UserID	RegistrationTime	AchievementLevel	NormalGrade	DistinctionGrade
1000	7 Oct '14 19:00	normal	84	47
2000	9 Oct '14 01:05	distinction	97	98
3000	10 Oct '14 20:00	normal	82	49
4000	10 Nov '14 13:36	distinction	94	96

Sample data about Clickstream Event

Id	UserID	EventType	Timestamp	LectureID
25000	1000	pageview action	10 Nov '14 16:01	103
25001	1000	video action	10 Nov '14 16:03	103
25002	1000	pageview action	10 Nov '14 16:42	104
25003	3000	pageview action	11 Nov '14 02:05	103
25004	2000	pageview action	11 Nov '14 02:15	103

Sample Data about Lecture

Id	Title	OpenTime	SectionID
103	Lecture 1.1: [. . .]	3 Nov '14 00:00	16
104	Lecture 1.2: [. . .]	3 Nov '14 00:00	16
105	Lecture 1.3: [. . .]	3 Nov '14 00:00	16
106	Lecture 2.1: [. . .]	10 Nov '14 00:00	16

We subdivide students into separate groups based on the assumption that similar group of students exhibit common behaviors. The first criteria for grouping is the type of certificate students enroll for. In order to acquire a signature track certificate, one is required to pay a fee and this motivation can translate into the exhibited level of commitment to learning. The second criteria is the achievement level or final grade. By clustering students according to their performance, we can highlight common characteristics and important inherent patterns. A detailed analysis can be found in [9]. For illustrative purposes, we only provide selective displays in this paper.

3.1 Visualising Viewing Behavior

We make use of the dotted chart in order to visualize the path followed by students while viewing videos. This provides a broad representation of students' watching behavior throughout the course.

In Figure 1, the dotted chart depicts the viewing behavior for all the students having registered for the MOOC focusing on when and how they watch videos. The x-axis depicts the time expressed in weeks, while the y-axis represents students. Seven different colors represent different events at a given time as carried

Table 3: Example of mapped cases and events for event log based on the data in Table 2

Sample details of Cases

CaseID	RegistrationTime	AchievementLevel	NormalGrade	DistinctionGrade
1000	7 Oct '14 19:00	normal	84	47
2000	9 Oct '14 01:05	distinction	97	98
1000	7 Oct '14 19:00	normal	84	47
2000	9 Oct '14 01:05	distinction	97	98

Sample details of Events

CaseID	Activity	Resource	Timestamp	AchievementLevel
1000	Lecture 1.1: [...]	1000	10 Nov '14 16:01	normal
1000	Lecture 1.2: [...]	1000	10 Nov '14 16:42	normal
2000	Lecture 1.1: [...]	2000	11 Nov '14 02:15	distinction

by students. The white dots show the timing when students viewed Miscellaneous videos (two videos on course background and introduction to tools). All videos for Week 1 are depicted with the blue dots, green dots represent videos for Week 2, gray dots show the distribution for videos in Week 3, all yellow dots show lecture views for Week 4, Week 5 videos are seen in red while Week 6 lecture videos are depicted by dark green dots.

Looking at this visualisation, we can observe that:

- A significant number of students drop out throughout the duration of the course. This can trigger further investigations to identify the deriving cause, i.e, follow-up emails, and take appropriate actions.
- Many stop watching after the first week but about 50% of students drop out after the second week of the course. Some actually even quit after watching the introductory videos (a handful of them).
- Not all students watch the videos in sequence. Although all of them watch Week 1 before watching Week 2, Figure 1 also demonstrates that even towards the end of the course, many still watch Week 1 and go back and forth. This also indicates that some videos are watched repeatedly and that a number of students progressively join the course later than the starting date.

In order to get detailed insights of this trend, we can also group the students into subgroups based on their respective profiles. We make this classification based on their final performance (distinction, normal and fail) and the type of the certificate they sign up for (signature or non-signature track). A detailed visualization for these respective groups can be found in [9]. We only consider distinction students on signature track as an illustration in Figure 2.

In Figure 2, these students follow a sequential pattern as they watch the videos. Some join a little late at Week 2 or Week 3, but the general trend re-

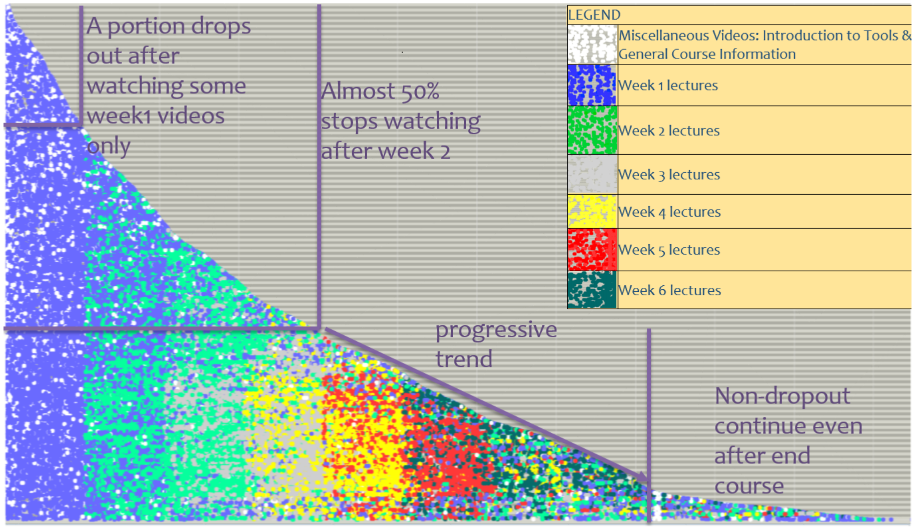


Fig. 1: Dotted Chart depicting a general viewing behavior throughout the duration of the MOOC

mains that most of them watch videos sequentially as they are made available. This can be seen by looking at the demarcation imposed by respective lecture videos colors. This is also captured by the process models as depicted in Figure 3. Successful students follow videos sequentially with orderly loops while unsuccessfull students appear to be volatile and unpredictable in their watching pattern. In the next section, we look at the respective process models depicting both successful and unsuccessful students' learning paths.

3.2 Process Discovery

Process discovery entails learning a process model from the event log. One can make use of an event log as an input to a number of process mining algorithms in order to visualize and enact the real behavior (sequential steps) of students. We used the fuzzy miner to mine our dataset. Rather than showing all the students' process models, we consider for illustrative purposes 2 extremes: the distinction students on signature track and failing students not on signature track. The resulting models are displayed in Figure 3.

The models in Figure 3 indicate that distinction students tend to have a more structured learning process, with a single path where possible loops are highlighted. On the contrary, the failing students follow a very unstructured learning process that exemplifies the volatitlity and unpredictability of their learning patterns. Although the fuzzy miner only shows the most dominant behavior, Figure 3 still shows that there are many alternative paths.

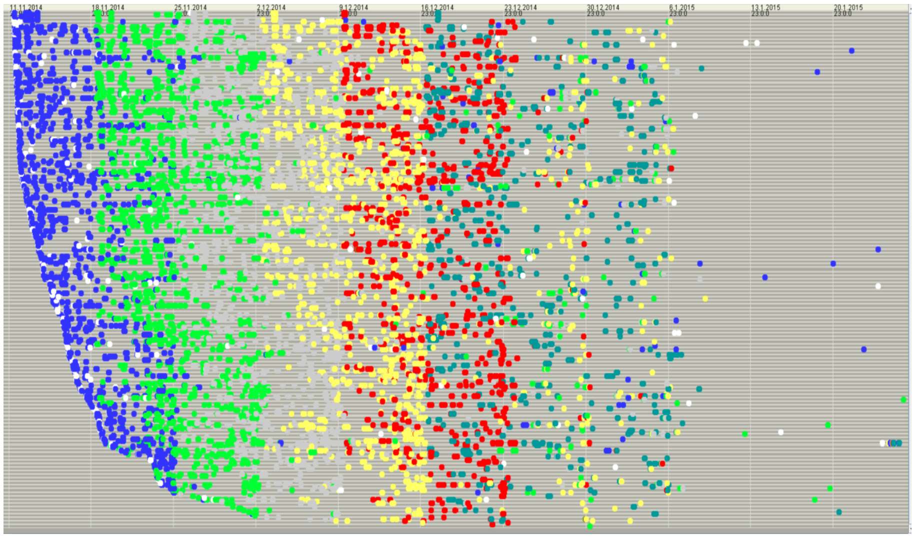


Fig. 2: Dotted Chart for Distinction Students on Signature Track.

4 Conformance Checking and Learning Behavior

Conformance checking can be used to uncover and quantify behavioral differences exhibited by different groups of students. Here, we use the alignment-based conformance checking approach [2, 3]. There are 2 critical aspects of students' behavior that we observe from the results of the conformance checking: *watch status* and *viewing habit*. With *watch status*, we aim at determining the sequence according to which each video is played, while the *viewing habit* defines the interval time between successive videos. These insights were obtained after performing conformance checking [1, 2].

Making an assumption that all students follow the course in sequence, we designed a model to represent this hypothesis. This idealised model, also called normative model, is depicted in Figure 4. It is an aggregated version of the real Business Process Modeling Notation (BPMN) model that shows only succession and flow between videos from Weeks 1 to 6. The main reason for not showng all videos in a chain is the high number of videos in the MOOC. With over 60 videos, the model would not be readable in this paper. Instead, the model used in the experiment specifies the first lecture in the series "Lecture 1.1: Data Science and Big Data (17 min.)" as the first task and the last lecture "Lecture 6.9: Data Science in Action (9 min.)" as the last task in the model. We also note the lectures we skipped due to space constraints (Lecture 1.3 to Lecture 3.8 and Lecture 4.2 to Lecture 6.8).

Following Figure 4, we performed alignment-based conformance checking [2] and a detailed description of this analysis is provided in [10]. By exploring the alignment details, we can thus analyze students' learning behaviors. Specifically,

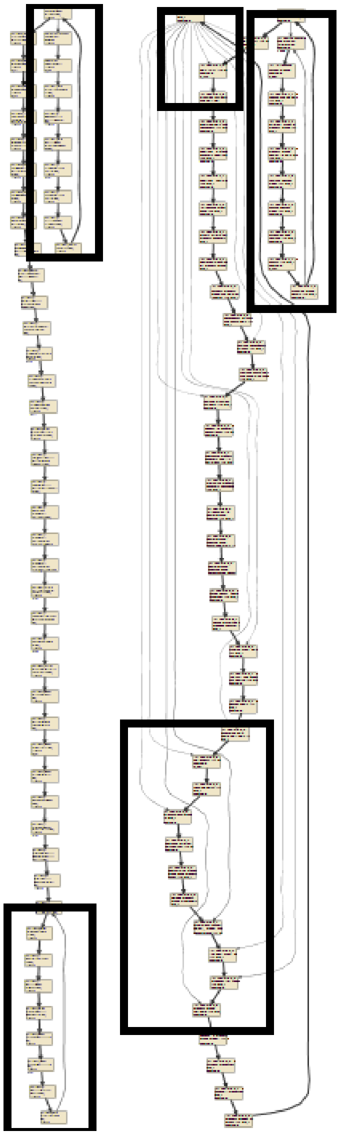


Fig. 3: Process Models for Signature Track Distinction Students with possible "loopbacks" vs. Non-Signature Track Fail students with "loopbacks and deviations"

we can visualize details about the overall *watch status* and *viewing habits* in sections 4.1 and 4.2.

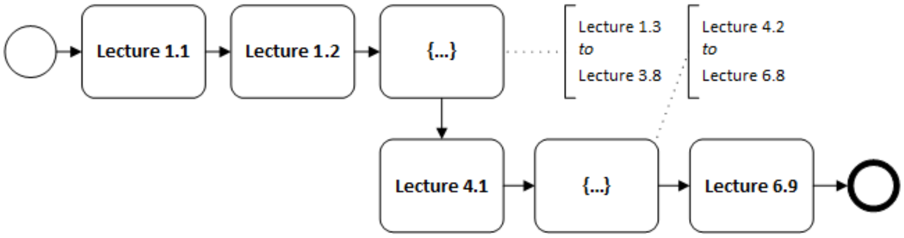
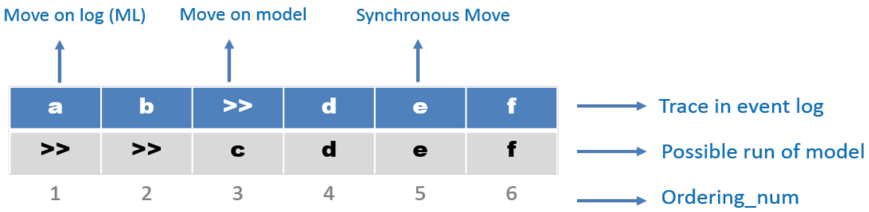


Fig. 4: BPMN Model for Sequential viewing of videos from Lecture 1.1 in Week 1 to Lecture 6.9 in Week 6

4.1 Video Watch Status

In order to label a video status, we consider moves that are generated by conformance alignment as seen in Figure 5. There are 3 types of moves that can be generated as a result. A move on log occurs when the task is found in the log only, a move on model occurs when it is only found in the model, and a synchronous move occurs in both the log and model [2]. Hence, looking at these three moves, we define the *watch status* as follows:



a	b	>>	d	e	f
>>	>>	c	d	e	f
1	2	3	4	5	6

Fig. 5: Conformance Alignment moves

```
SET Watch Status =
CASE WHEN move = ‘synchronous’ then ‘WatchedRegularly’
        WHEN move = ‘modelOnly’ then ‘NotWatched’
        WHEN move = ‘logOnly’ then
                CASE WHEN ordering_num in model < ordering_num in log
                              then ‘WatchedEarly’
                            ELSE ‘WatchedLate’
                  END
```

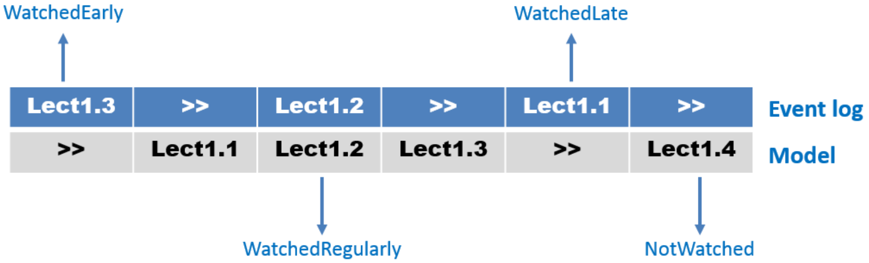


Fig. 6: Description of videos *watch status*

END

We illustrate with a single possible run of log with 4 transitions (lectures): Lect1.1, Lect1.2, Lect1.3 and Lect1.4. We also consider an event log with trace $\langle$*Lect1.3, Lect1.2, Lect1.1*$\rangle$. With conformance alignments, we can identify the videos *watch status* as depicted in Figure 6. The overall videos status for signature track students for the duration of the course is presented in Figure 7. Detailed results for other subgroups of students can be found in [10].

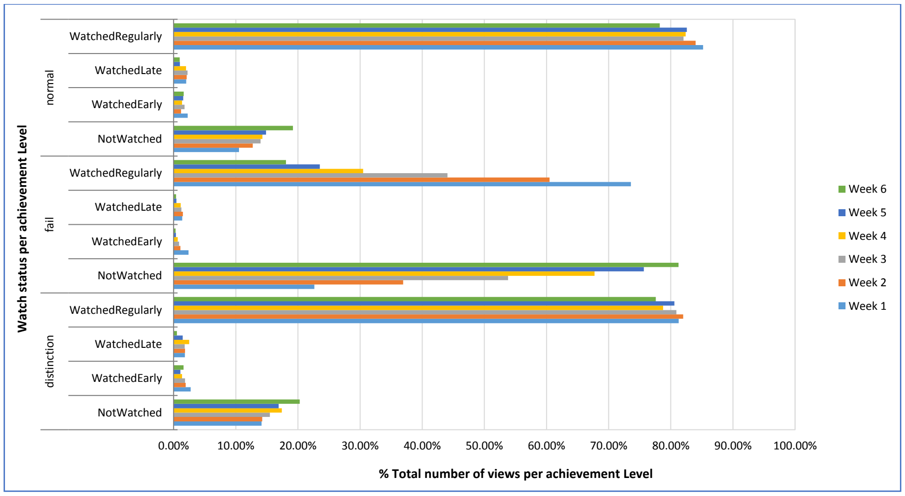


Fig. 7: Overall *watch status* per week for distinction, fail and normal students

Figure 7 indicates that successful students are consistent with watching videos in sequence. The graph points out that the *watch status WatchedRegularly* is dominant for successful students while there is a slight progression for *NotWatched*. Unsuccessful students in most parts progressively stop watching videos from Week 1 and the trend can be observed increasing until Week 6.

4.2 Viewing Habit

The *viewing habit* describes the time commitment in the students' learning behavior. It depends on the time at which two successive videos are opened. In order to define these habits, we count the number of minutes between open times for successive videos and specify the thresholds as follows:

```
SET Viewing Habit =
CASE WHEN interval $\leq$ 30 then 'InBatch'
          WHEN interval $\leq$ 60 then 'After30min'
          WHEN interval $\leq$ 120 then 'Hourly'
          WHEN interval $\leq$ 720 then 'Halfdaily'
          WHEN interval $\leq$ 1440 then 'Daily'
          WHEN interval $\leq$ 10080 then 'Weekly'
          ELSE 'Skipped'
END
```

Figure 8 shows a representation of students' habits over 6 weeks. There is a clear indication of the impact of *viewing habit* on performance and students' final grades. The most committed students, who watch mostly in batch appear to be more successful than the rest.

The opposite trend is observed with regards to unsuccessful students who increasingly skip videos. As the MOOC starts, some of these students are devoted to watching but as time progresses, they stop watching certain videos and this shows accross the board for all unsuccessful students. Moreover, unsuccessful students' behavior pertaining to watching in batch progressively decreases as the weeks go by. The more videos were watched in batch in Week 1, the less they are in Week 6.

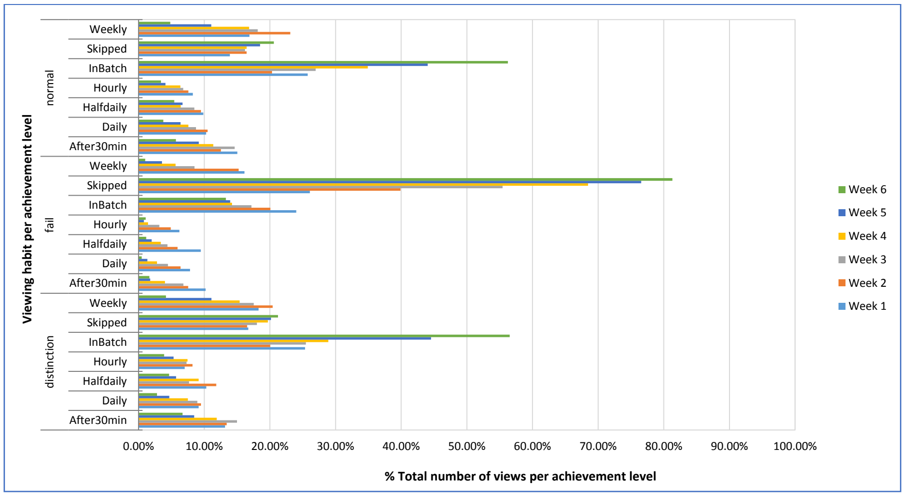


Fig. 8: *Viewing habits* per week for distinction, fail and normal students

4.3 Viewing Habit vs. Watch Status

It is also interesting to visualize the correlation between *viewing habit* and *watch status*. Some of the questions we might try to answer are: “Are students who watch videos in batch watching videos sequentially?”, “Is there a link between both *watch status* and *viewing habit*?”. In Figure 9, we observe that students who study in batch, mostly watch videos regularly (in sequence) than those who skip videos.

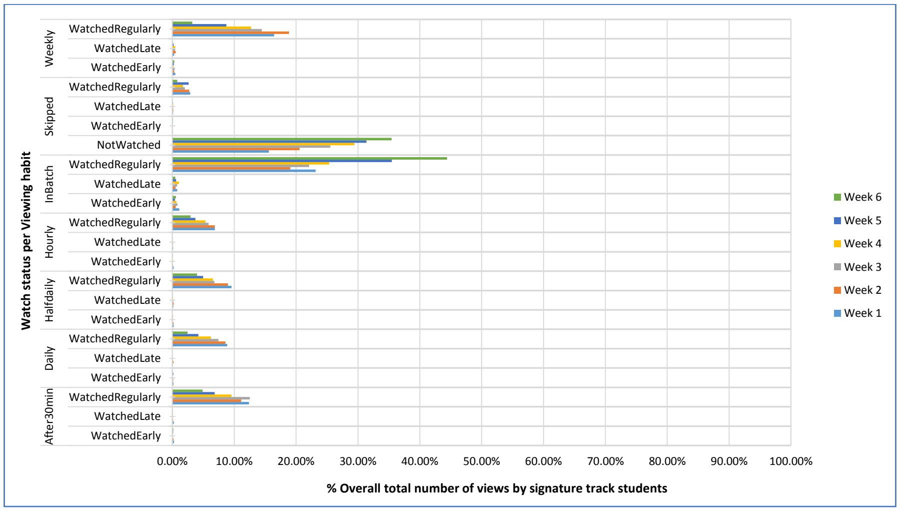


Fig. 9: *Watch Status* versus *Viewing Habit*

Figure 9 shows a relationship between the way people watch videos and the interval of time between successive videos. It indicates that when students watch videos in batch, they are more likely to follow a proper sequential pattern indicated by status *WatchedRegularly*.

5 Measuring Correlation Between Learning Behavior and Performance

Measuring the correlation between learning behavior and performance can help to quantify the observations made in sections 3 and 4. We consider three different statistical measures of correlation in order to determine the level of the relationship between how students watch lecture videos (behavior) and their performance (final grades). We calculate the Pearson’s, Kendall’s and Spearman’s correlation coefficients [6].

The Pearson’s coefficient determines the degree of a relationship or a linear correlation between two attributes, i.e. students’ viewing behavior and final grades [6] while the Kendall's and Spearman's correlation coefficients determine the degree of a relationship between two attributes based on ranked data [6]. These coefficients are between -1 and +1 indicating the degree of the correlation. The plus (+) and minus (-) signs also indicate whether it is a negative or a positive correlation.

We express, for this analysis, the learning behavior by the trace fitness value and label it as *Viewtrend* in Figure 10. We then compute the correlation coefficients between *Viewtrend* and students' final grades (normal and distinction grades). The values of *Viewtrend* are between 0 and 1 indicating the scale of students' learning behavior, while the final grades span from 0 to 100. Figure 10 shows two graph matrices for both signature and non-signature track students. In each matrix, three graphs (2 histograms and 1 scatterplot) are produced. The histogram at the intersection of the same attribute, i.e NormalGrade x NormalGrade, captures that attribute's distribution for the considered population, i.e signature-track students, while the scatterplots at the intersection of different attributes show their correlation distribution. On each scatterplot, the correlation coefficients are given in red, i.e 0.39/0.26/0.37, representing respectively the Pearson's, Kendall's and Spearman's correlation coefficients. We observe a moderate and positive correlation as indicated by the values of the coefficients in Figure 10. Particularly, we observe that the Pearson's correlation coefficient has the highest value for both signature and non-signature track students given by values 0.39 in Figure 10a and 0.55 10b.

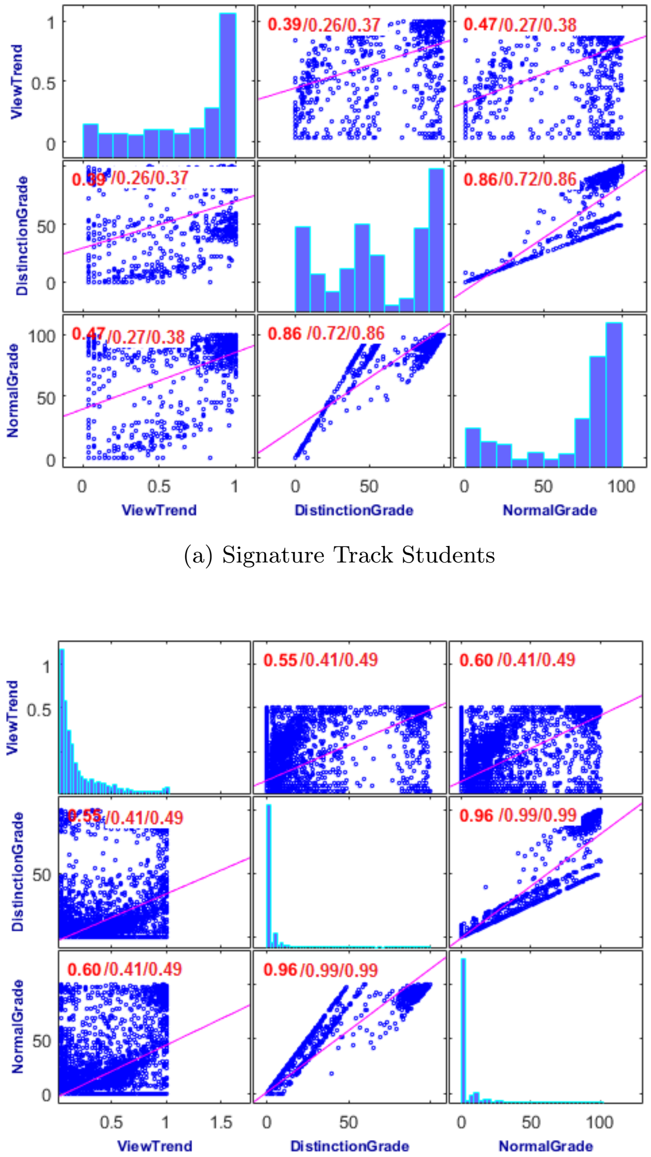
These results are critical as they indicate the existence a positive relationship between the way people watch videos and the outcome of their performance. Nevertheless, the values of the correlation coefficients (between 0.26 and 0.60) also indicate that additional factors such as students' background, focus level, previous content knowledge, IQ level etc. can be contributing factors to the overall performance.

6 Conclusion

Learning Analytics (LA) [5, 7] promises to provide insights from educational data. In addition to current LA work based mostly on traditional data mining techniques [5], this paper proposes to use process mining in order to provide insightful analysis based on the actual behavior of students.

Taking our Coursera MOOC as a case study, we show the added value of process mining on MOOC data. Our results demonstrated that the way students watch videos as well as the interval between successive watched videos have a direct impact on their performance. Results indicate that successful students follow a sequentially-structured watching pattern while unsuccessful students are unpredictable and watch videos in a less structured way.

Moreover, students' learning behavior can be described from two dimensions: *watch status* and *viewing habit* as described in section 4. The results indicate that



(a) Signature Track Students

(b) Non-Signature Track Students

Fig. 10: Pearson's/Kendall's/Spearman's Correlation Coefficients Matrix for Students' *Viewtrend* and Final Grades(Normal and Disinction)

in general, students' *viewing habit*s are determined by the time between successive videos while the *watch status* is determined by the conformance alignments. Our results identified that students who watch videos regularly and in batch are more likely to perform better than those who skip videos or procrastinate in watching videos.

Finally, we calculated three statistical measures of correlation considering the Pearson's, Kendall's and Spearman's correlation coefficients. The calculated coefficients ranged between 0.26 and 0.60 confirming the existence of a positive correlation between learning behavior and performance in a MOOC as seen in Figure 10.

In the future, we aim at conducting additional experiments using other process mining techniques described in [1] and analyze other paradigms of LA on MOOCs.

References

1. van der Aalst, W.M.P.: Process Mining - Discovery, Conformance and Enhancement of Business Processes. Springer (2011)
2. van der Aalst, W.M.P., Adriansyah, A., van Dongen, B.F.: Replaying history on process models for conformance checking and performance analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(2), 182–192 (2012)
3. Adriansyah, A., Munoz-Gama, J., Carmona, J., van Dongen, B.F., van der Aalst, W.M.P.: Alignment based precision checking. In: Business Process Management Workshops. pp. 137–149. Springer (2013)
4. Arnold, K.E., Pistilli, M.D.: Course signals at Purdue: using learning analytics to increase student success. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. pp. 267–270. ACM (2012)
5. Baker, R.S., Inventado, P.S.: Educational data mining and learning analytics. In: Learning Analytics, pp. 61–75. Springer (2014)
6. Chok, N.S.: Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data. Ph.D. thesis, University of Pittsburgh (2010)
7. Ferguson, R.: Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning 4(5-6), 304–317 (2012)
8. 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)
9. Mukala, P., Buijs, J.C.A.M., van der Aalst, W.M.P.: Exploring students' learning behaviour in moocs using process mining techniques. Tech. rep., Eindhoven University of Technology, BPM Center Report BPM-15-10, BPMcenter.org (2015)
10. Mukala, P., Buijs, J.C.A.M., van der Aalst, W.M.P.: Uncovering learning patterns in a mooc through conformance alignments. Tech. rep., Eindhoven University of Technology, BPM Center Report BPM-15-09, BPMcenter.org (2015)
11. Siemens, G., de Baker, R.S.J.: Learning analytics and educational data mining: towards communication and collaboration. In: Second International Conference on Learning Analytics and Knowledge, LAK 2012, Vancouver, BC, Canada, April 29 - May 02, 2012. pp. 252–254 (2012)