How well can completion of online courses be predicted using binary logistic regression?

Andersson U, Arvemo T and Gellerstedt M

University West
Sweden

Abstract. This article uses binary logistic regression to create models for predicting course performance. The data used is the data-trail left by students activities on a discussion forum while attending an online course. The purpose of the study is to evalute how well models based on binary logistic regression can be used to predict course completion. Three sets of data was used for this. One set collected at the end of the course, one collected after 75% of the course and one set collected after half the course. The result of the study says that it's possible to design models with an accuracy of between 70% and 80% using these methods, regardless of what time is used.

Key words: online education, statistics, course completion, online forums, educational data mining

1 Background

Online courses, where content, communication and assignments are mediated via the Internet or similar channels are becoming more and more common. While they have their advantages, they also suffer from a low completion rate compared to more traditional courses. Completion rates normally is around 30% to 40 % in these courses [1], and this also corresponds with the experience of the authors. The numbers for more traditional courses on campus is from 75% to 95%.

While not all students can be helped, there is a group of students with the time and motivation to fullfil their studies, but for various reasons they run into problems. These students are often helped by intervention and support from the teacher [2] and they are described as ”Low hanging fruit” by [3], meaning that a small effort put into helping these students has a high return in regards to students completing the course.

However, in an online environment, the teacher have a harder time identifying the signals of a student that is slipping behind [4]. Instead of relying on visual and other cues that the teacher can perceive in the contact with the students other solutions have to be found. One way is to utilise the trail of data left by the students when interacting with each other and the course material via the platform the course is using[4].

Several disciplines and academic areas have emerged, studying the use of that data trail and how it can be used to predict different aspects of student performance. In a literary review, Sin finds 8 different areas of interest for these kinds of studies [5]. Of these, behaviour detection is the most common, Performance prediction (predicting the grade of the student) is in the middle and Attrition Risk (the risk that the student fails or dropps out of the course) is the least studied area.

The different disciplines all use some kind of big data-approach, where all the data produced by the students are used for the analysis. The difference is mostly based on the goals and focus on the usage of this data. Two of the larger disciplines is Educational Data Mining and Learning Analytics [4].

Educational Data Mining is more focused on the technical challenge, how to interpret and get valuable information from data gathered from the data trail left by the students. *Learning analytics* on the other hand is about how to optimise the learning opportunities.

While there is a slight difference in focus and goal between these two, there is also a significant overlap in these two areas of research. At their core, they use many similar techniques and methods, and in the end their goal is the same.

Many of the models and studies made using educational data mining has focuses solely on the data trail left by the students, and hasn't taken learning theories into concideration. While that is a common approach when talking about big data [6] it has also been criticised, and Gasevic and others emphasise the need to use learning theories when performing educational data mining [7]

1.1 Engagement and learning

When looking at different factors that correlates with learning, student engagement and activity is often cited as the most decisive factor. [8] [9] In early studies, this was described as "handing in the assignments", but was later on expanded to things like helping and interacting with fellow students and engaging in areas related to, but not explicitly within the boundaries of the course. [9]

1.2 Factors related to student performance

We know that it's important to identify and support students that risk falling behind in their studies. We also know that, overall, student engagement and activity is important factors that determine student success. However, applying this knowledge to online courses and educational data mining should be done carefully.

There are a number of factors influencing student success within a course. Some of those are more related to areas like student background, previous experiences with academic studies or equipment and internet connection [2]. To handle this, Conrad and Donaldson, in their book Engaging the online learner [10], suggests a survey where the online student performs a self evaluation regarding his or her readiness for online learning.

Several studies emphasise the importance of student engagement and activity These factors are among the most significant for predicting how well the student performs in a course. [11]

One study regarding student activity and course performance was performed in 2004 by Webb et al. Their findings about forum activity and learning outcomes say, among other things, that 60% of the students that didn't participate in forum discussions failed the course. Also, just looking at the two factors of accessing and posting to the forum they found that between 10% and 20% of the grade variance could be accounted for by using these variables as predictors. [11]

1.3 Previous study

This study expands a previous study by Andersson et al[12]. The purpose of that study was to examine a number of factors, chosen to fit two conditions. They should be based on the theories of activity and engagement (including helping others and engaging in extra-curricular discussions), and they should be simple to collect and analyse. The following factors were examined: The number of posts made, the average length of the posts, the number of threads started, the number of replies made to others, the number of non compulsory posts made, the number of non-compulsory threads started, the average number of posts made / day of activity, the number of active days and the average length of chains of consecutive days of activity.

That study examined how well each factor correlated with the grades of the course, one by one. Overall, all the factors had a correlation with the grades one way or another. However, the factors were studied one by one, and at the end of the course. This study aims to expand on that study in two ways: Use a binary logistic regression, thereby examining how different factors work together. Consider data not only from the end of the course, but also from times while the course is still running. This is crucial if we want to be able to detect students while the course is still running.

1.4 Purpose

The strategic purpose of this study is to form a foundation for an automated system to monitor students on online courses in order to identify those students that run the risk of failing the course. This system would work from what we know about student performance, as described below.

We know that student performance is linked to student engagement and activity, both in traditional courses and in online courses.

The purpose of this study is to evaluate how well course completion can be predicted using automatically collected data on online activity.

In this study, we limited the factors evaluated to factors that are simple to collect and analyse by automatic means, and that are considered to be indicators of student engagement.

A secondary purpose is to compare how the importance of different online activities change during the span of the course.

2 Method

The study was performed on an online class in 3d-graphics. All communication between students and teacher, and between students, was done via an online forum using the phpBB-platform (an open source framework for building online discussion forums). All the assignments on the course were posted on a compulsory sub-section of the forum, but there were also sections of the forum that had a non-compulsory purpose. In these sections the students could ask for help, discuss things with each other etc.

For each of the four assignments, students were required to post at least one report themselves (including pictures), and after the deadline of the assignment, also give feedback to a peer student. In other words, in order to pass the minimum requirements for the course, at least 8 posts should be done, and 4 threads should be started.

The grades possible on this course was Fail, Pass and Pass with Distinction.

The number of active students on the course was 66. Out of these 33 failed the course, 18 passed the course, and 15 passed with distinction.

After the conclusion of the course, data from the forum was retrieved by data scraping using Python scripts.

The purpose of the study was to collect numerical data, thereby making it a quantitative study. The data was grouped based on the grades of the students, and then the data inside the groups was analysed. The students were informed about the study, but other than that the study was purely observational.

2.1 Data

The variables used in this study builds upon the study made by Andersson et al [12]. The main difference from that study is that the varibales are divided into two groups. The first group contains variables related to the numbers of posts made and the frequency of posting etc. In addition to the variables used in that study, this study adds two more variables in Number of medium length runs, and number of long length runs. One variable was also removed from the variables used in the previous study (average number of days between posts) because it didn't correlate with either the grades or if the students failed or passed the course.

The other group is just a differentiation of the Length-variable. Instead of just looking at the average length of the posts each student make, this variable is split into sub-variables depending on the type of post.

Number and frequency of posts In the list below, the name used in the upcoming tables for each predictor is written in the parenthesis.

Number of posts (Posts)
: The total amount of posts made on the forum during the course. At least 8 in order to pass the course

Number of replies (Replies)
This is the number of replies made to other students posts. At least 4 is required to pass the course, and more than that is a good indicator of an engaged student.

Number of non-compulsory posts (NC)
This is the number of posts made on the forum outside of the compulsory areas of the course. This corresponds with the view that student engagement is shown by activities outside of the minimum requirements of the course

Number of threads started (Threads)
The number of threads started on the course. At least 4 in order to pass the course

Number of non-compulsory threads started (NCThreads)
This is the number of threads started outside of the compulsory areas of the course.

Batch posting (avgBatch)
The opposite of the number of active days, this measures how many posts that are made each day of activity.

Chains of posting (avgChain)
Another variable that shows the pattern of posting. The value of this parameter is the length of chains of consecutive activity (every or every other day) on the forum.

number of medium length runs (medRuns)
This is the number of chains (see parameter 9) that runs for 4 or 5 days.

number of long length runs (longRuns
This is the number of chains (see parameter 9) that runs for more than 5 days

Length based variables When analysing length of posts, the posts where divided into several group.

Reply or ThreadStart
Posts where divided into whether they are the first post in a thread, or a reply within an already existing thread

Own or Other
Only applicable for replies, this grouping is whether a reply is to someone elses thread or to a thread the poster started earlier

Comp or NC
This tells whether the posts were made on the compulsory part of the forum or on the non-compulsory parts.

This different groupings were then combined, so that the length of Threadstarts in the compulsory part of the forum could be studied, as well as replies to other students threads in the non-compulsory part of the forum etc. Noticable groupings here are for instance posts that start threads in the compulsory parts of the forums. These are the students actual assignments.

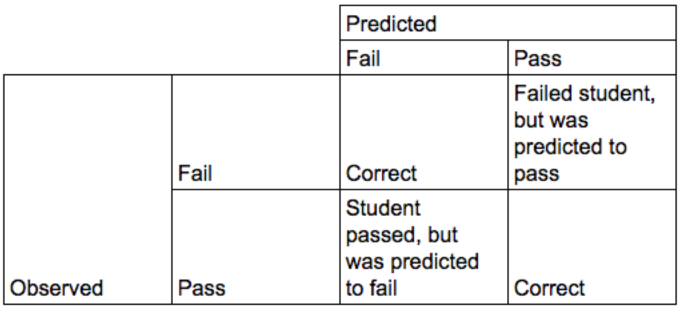
2.2 Binary Logistic Regression

Binary logistic regression is a statistical method that can be used for modelling the probability for a dependent variable to belong to one of two states. In this case, these two states are Pass and Fail. To estimate the model a set of observations including both the dependent and independent variables are used. When the model is estimated we can use the model to predict the probability of an observation of the dependent variable being in one of the categories given the values of the independent variables. [13]

In this particular study, each observation is the data trail left by each students online activities. As mentioned above, the two states are Fail and Pass.

This means that we can, for instance, create a statistical model from variables like Number of Posts, Length of posts and the number of replies made. When the model is created, we can enter new values for those variables, and the model gives us the probability for a student corresponding to those values to fail or pass the course.

One way to evaluate a model, and the one used in this study, is to re-enter the data used to build the model. If the probability for each observation is equal to, or more than, a certain value (often 0.5) that individual is said to belong to the second, or higher, state (in this case: Passed). For each individual, there are four possible outcomes, as per the table below



		Predicted	
		Fail	Pass
Observed	Fail	Correct	Failed student, but was predicted to pass
	Pass	Student passed, but was predicted to fail	Correct

Fig. 1. Active days

The accuracy of a modell is the proportion of correct predictions of both kinds.[13]. In this study, the two types of wrong predictions will be treated differently. This is because the aim of the study is to identify students who is at risk of failing the course. Therefore, we want the number of Failed students predicted to pass (upper right corner in figure 1 above) to be as low as possible. However, a higher number of passed students that are predicted to fail are acceptable.

2.3 Big-data approach

Instead of designing a small number of well thought through models which where then tested, a big-data approach was adopted. Python, and the libraries numPy, Pandas and statsModels were used for all analysis, and this made it possible to write scripts to test a large number of models. Three lists of possible models were created through the help of Python. One list contained all possible combinations of factors (up to three factors), using the factors belonging to the first category mentioned above. The second list consists of all combinations of the length based factors (up to three), and the third list consists of all combinations together (again, up to three factors). In order to get an idea of which models had the best predictive ability at different times during the course, this process was repeated using data from day 35 (of 70) of the course, day 53 of the course and at full time, that is day 70 of the course.

3 Results

The tables 1, 2 and 3 show the models at the three different times, at 35 days, at 53 days and at 70 days

Each section of the table lists the five most accurate models from each list of factors. The upper section lists models created using only post- and frequency based factors, the middle section lists models created using only length based factors and the lower section lists models created using all factors.

4 Summary of results

Overall, the accuracy (the number of cases the model manages to guess correctly) is around 0.65 to 0.7 for the 35-day model, from 0.7 to 0.76 for the 53 day model and around the 0.8-mark for the 70 day model.

For all three time periods the model based on number of posts, number of non-compulsory posts and the average batch (that is, number of posts made each day) had the highest accuracy among the models based factors describing posts and frequency. On position 2-5 in the lists, the factors vary slightly between the different times. Noticeable is that the average batch-factor appears in all models except for one.

In the earlier models, number of posts also appear frequently, whereas number of threads, replies and non-compulsory posts start to make a difference in the models created from data later in the course.

Looking at the length-based-models, there is a bigger variety among the factors. Also, apart from the models created by data from day 35, the accuracy is very similar between all length-based models.

Variables	Acc	Missed Fails	Missed Pass
Posts NC avgBatch	0.65	15	8
Posts avgBatch avgRun	0.65	12	11
Posts avgBatch	0.64	9	15
Posts avgBatch MedRuns	0.64	9	15
Posts avgBatch NCThreads	0.64	11	13
LengthReplyToOwnComp LengthThreadStart Length	0.7	11	9
LengthNCReplies LengthReplyToOwn LengthCompThreadStart	0.65	8	15
LengthReplyToOwnComp LengthNCReplies LengthCompThreadStart	0.64	6	18
LengthNC LengthCompThreadStart Length	0.64	19	5
LengthNCReplies LengthReplyToOwn LengthThreadStart	0.64	8	16
LengthReplyToOwnComp LengthThreadStart Length	0.7	11	9
Posts LengthCompThreadStart avgBatch	0.7	11	9
LengthThreadStart Posts avgBatch	0.68	12	9
LengthNC Posts avgBatch	0.67	12	10
LengthNCReplies LengthReplyToOwn avgBatch	0.67	7	15

Table 1. Day 35

Variables	Acc	Missed Fails	Missed Pass
NC Posts avgBatch	0.73	13	5
Threads Replies avgBatch	0.71	14	5
NC Replies avgBatch	0.7	11	9
Threads Posts Replies	0.7	8	12
Posts avgBatch longRuns	0.7	9	11
LengthReplyToOwnComp LengthCompThreadStart	0.76	5	11
LengthReplyToOwnComp LengthThreadStart	0.76	5	11
LengthReplyToOwnComp LengthReplyToOwn LengthCompThreadStart	0.76	5	11
LengthReplyToOwnComp LengthReplyToOwn LengthThreadStart	0.76	5	11
LengthReplyToOwnComp Length	0.74	6	11
LengthReplyToOwnComp LengthCompThreadStart	0.76	5	11
LengthReplyToOwnComp LengthThreadStart	0.76	5	11
NC LengthReplyToOwnComp LengthCompThreadStart	0.76	5	11
NC LengthReplyToOwnComp LengthThreadStart	0.76	5	11
LengthReplyToOwnComp Threads LengthThreadStart	0.76	5	11

Table 2. Day 53

Comparing the different groups over time, the models based on length are slightly more accurate in the beginning and middle part of the course, whereas the models based on posts and frequency are more accurate at the end of the course.

In addition to looking at the accuracy as a whole, it's also interesting to note that there is a substantial difference between different models in their ability to predict failures or passes. For instance, the models created from length-based

Variables	Acc	Missed Fails	Missed Pass
NC Posts avgBatch	0.82	10	2
Replies avgBatch	0.8	10	3
avgRun Replies avgBatch	0.8	10	3
Replies MedRuns avgBatch	0.8	9	4
Replies Posts avgBatch	0.8	10	3
LengthThreadStart LengthReplyToOwnComp	0.77	6	9
LengthReplyToOwnComp Length	0.77	6	9
LengthCompReplies LengthThreadStart LengthReplyToOwn	0.77	5	10
LengthThreadStart LengthReplyToOwnComp LengthReplyToOwn	0.77	6	9
LengthReplyToOwnNC LengthReplyToOwnComp LengthCompThreadStart	0.77	8	7
LengthNC Posts avgBatch	0.83	7	4
NC Posts avgBatch	0.82	10	2
Replies avgBatch LengthNCThreadStart	0.82	9	3
Replies avgBatch	0.8	10	3
avgRun Replies avgBatch	0.8	10	3

Table 3. Day 70

data at day 53 and at day 70 show a difference in how many failed students it failed to predict versus how many passed students it failed to predict. In both these cases, it was more accurate in predicting failures than passes.

Overall, the results seem to indicate that after half the course, it's possible to identify about two thirds of the students who run the risk of failing the class. After 75% of the course, the number of correct predictions of failed students increases to 70-80%.

Types of failed predictions The binary logistic regression model does not just give a binary answer whether a student is predicted to fail or pass the course. Instead, it gives a probability estimate of the outcome. As well as talking about the number of failed estimates, it is also of some interest to talk about the degree of failure. In all models there was students just below the 0.5-mark that passed the course although they were flagged as predicted failures, just as the other way around.

In addition to that, there are also outliers that show a very high probability for passing the course, but still fails, and the other way around.

5 Conclusion, Discussion and further studies

The main purpose of this study is to evaluate how well course completion can be predicted using automatically collected data on online activity. While the exact percentages may not be repeatable for other courses, it seems plausible that it's possible to predict course performance using data collected in this way. For instance, the overall accuracy of the best models at day 53 (or three quarters into the course) is around 75%. Some of these models are, in addition to this, even better at identifying students who risk not completing the course.

The secondary purpose of the study was to see how the importance of different variables change over the duration of the course. Overall, it seems that in length-based variables, or a combination of length based and other variables that are not as time dependent (like avgBatch) are more important. The longer the course progresses, the more important variables like Posts or NC (Non-compulsory posts) become.

One issue that is raised is the issue of which students we want to capture. While the natural assumption would be to find the parameters for students who fail the class, that doesn't necessarily be the only way to work. it is easy to see that students who barely pass the course, and students that barely fail it, very well could display the same behaviour. An in-depth study where the subgroups "Strong but Failing students" and "Weak, barely passing students" are compared to each other, and to students with stronger grades (in the case of the latter). This would require a finer evaluation of the students than what the grades would give, which in turn means that a close cooperation with the course examiner would be a prerequisite.

Different models had different success rates when predicting failures and passes. For instance, one model could be very good at identifying the students who were about to pass the course, but also included a number of failing students into the group of passing students. Another model could be very good at predicting which students that fail, but including passing students into the failing group.

The ideal would be to have a model that have a high accuracy for both groups (low number of wrong predictions for both groups), but if that's not possible, one has to consider the consequences for the two types of errors.

Let's say that there is a system using these kinds of models to predict and give notifications about student behaviour. A student that is wrongly predicted as a passing student might go unnoticed by the course teacher, and therefore not getting the support that is needed. On the other hand, passing students that are flagged as failing are brought to the teachers attention. After that, the teacher could look at the work of the student so far and make a judgement of his or her own to see if the students needed more support. In the case of a fully automated system, the result would be that a student that is on his or her way of passing the course gets an email or similar asking of support is needed.

While the latter could be annoying, overall it's a better outcome than what happens if a failing student is pointed out as a passing student. Overall, this means that in the choice between a model that is good at predicting passing students but not failing, and a model that does the opposite, the latter is preferred.

Outliers All models contained cases where the prediction of whether the student should fail or pass was clear. Either the prediction was very high (0.8 or higher) or very low (below the 0.2 mark), but the prediction was still wrong.

While further studies are required to see what happens here, there are still some plausible explanations.

For students with a high probability of success, this could mean that they for all purposes intended to finish the course, but circumstances changed. Maybe things changed at work (or they went from being unemployed to employed), family situation changed or similar.

The opposite is not as natural to explain. These would be those who pass the course, but act in ways similar to those who usually fail it. One group of students who might be found here are those who already know the subject, but also want the course credits for it. For them, it's not a case of being engaged in the course as such, but just finding the minimum amount of effort required to pass the course, thus getting the credits for what they already know.

5.1 Further studies

While this model seems to be able to predict the drop-out-rate of the students to an acceptable degree it is still only tested on this particular course. While there are some generic tendencies that seem to be applicable to any course, it's not something that should be taken for granted [7]

One thing that is not done in this study is to calculate the contribution and significance of the variables in the different models. This is something that could be done in future studies, in order to better evaluate the different models.

As suggested by Conrad et al [10], it is also important to take the individual skills and experiences of students into consideration. A follow up study should also do a self-evaluation survey, and map the parameters found in this study not only to the course outcomes, but also to their replies on that survey. For instance, an experienced online learner might be better to complete the course by just fulfilling the minimum requirements, and still be in no danger of failing the class, whereas for a more inexperienced student that might be indicators of someone on the verge of failing the class.

References

1. Clow, D.: MOOCs and the funnel of participation. In: the Third International Conference, New York, New York, USA, ACM Press (2013) 185–6
2. Rankin, D.: Predictors of Success for High School Students Enrolled in Online Courses in a Single District Program. PhD thesis (2013)
3. Zhang, H., Almeroth, K., Knight, A., Bulger, M.: Moodog: Tracking students' online learning activities. World Conference on Educational Media and Technology (2007)
4. Ferguson, R.: Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning **4**(5/6) (2012) 304–15
5. Sin, K., Muthu, L.: Application of big data in education data mining and learning analytics—A lterature review. ICTACT Journal on Soft Computing (2015)
6. Mayer-Schönberger, V., Cukier, K.: Big data: A revolution that will transform how we live, work, and think (2013)

7. Gašević, D., Dawson, S., Siemens, G.: Let's not forget: Learning analytics are about learning. TechTrends **59**(1) (2015) 64–71
8. Davies, J., Graff, M.: Performance in e-learning: online participation and student grades. British Journal of Educational Technology (2005)
9. Chapman, E.: Assessing Student Engagement Rates. ERIC Digest. (2003)
10. Conrad, R.M., Donaldson, J.A.: Engaging the online learner. 27th Annual Conference on Distance Teaching (2004)
11. Webb, E., Jones, A., Barker, P.: Using e-learning dialogues in higher education. Innovations in Education and Teaching International **41**(1) (2004)
12. Andersson, U., Gellerstedt, M., Arvemo, T.: Can measurements of online behavior predict course performance? In: IMCIC. (March 2016)
13. Samuels, M.L., Witmer, J.A., Schaffner, A.: Statistics for the Life Sciences. 5th edn. Pearson College Division, Harlow (2016)