Drop-out prediction from digital learning for retention

PROJECT PARTICIPANTS:
Bianca Clavio Christensen, Hendrik Knoche, Bastian Ilsø Hougaard, Ninna Vihrs, Jon Ram Bruun-Pedersen, Lise Busk Kofoed, Torben Tvedebrink, Poul Svante Eriksen, and Brian Møller.

Introduction

Dropout and retention in the first semesters of universities have been linked to a variety of factors in the literature including but not limited to high school performance, academic performance, demographics, motivation to study, and activities in virtual learning environments. From these factors, the University of Eindhoven University has achieved between 75-80% accuracy for predicting dropouts in [1] in their electrical engineering degrees that traditionally have high dropout rates (40%) in the first year. They use this information to advise students at risk and to allocate resources.

While much of the required information is being held by AAU, course and semester coordinators have no means of accessing this data nor obtain such an integrated analysis about which students are at high risk of failing and dropping out to direct resources at. Much of the necessary data is currently held by AAU. Medialogy is going to experiment with collecting relevant data regarding demographics and study motivation through the study verification test (studiestartsprøve, SSP). In an ongoing PBL development project (Improving Moodle for flipped classrooms to decrease dropouts) we introduced new interactive activities to AAU's Moodle for flipped classroom design to improve both student classroom participation and collect data performance on course participation and performance to facilitate analysis. These tools do not only serve the purpose of data collection but through their use can help students to reflect on their learning progress, e.g. learning inventories and other self--assessments that are very much in line with PBL practice.

Overview of project phases:

- Phase 1: Identifying study program specific success-factors from current data
- Phase 2: Identifying data that can improve prediction quality and start collecting these data
- Phase 3: Improve understanding of the risk group - trying to distinguish between students who chose the education by mistake vs. students who want to pursue the education but require help
- Phase 4: creating a web or application based prototype that can create analysis on demand using for exam outcome and drop- out prediction for different study directions

In Phase 1 through 3, we analysed data of the current Medialogy cohorts in R statistic program. In this process, we received data from QlikView and via data handlers (Ole Garsdal Hansen and Cristina Draghici). As a result of these phases, we sought to create and update the prediction models for each semester (ideally use all data available) with the purpose of providing individualized counselling based on significant predictors and make a better guess of the student's risk of dropping out compared to the current system. This information could be interesting for the study board, e.g. for changing enrolment or re-exam policies. To automate the data analysis process we needed direct access to the STADS database to create proper back-end development for the web application prototype in Phase 4. Due to administrative work of getting access to the STADS database and the GDPR, the full prototype ceases to function after this project period by the end of February 2019.

Identify study program specific success-factors and the risk group

We based the data analysis on one previous study on student dropouts in 2016 [2] that looked at first semester students in the Medialogy bachelor programme. While their findings from questionnaires, interviews, and study diary logs suggested that reasons for dropping out were quite diverse, they provided some evidence that the required skills and levels in mathematics and programming were higher than students initially expected to result in dropouts. To extend their analysis, we performed a prediction analysis based their findings and known factors in the retention literature. Through data analysis of student demographics and grades, we gained an understanding of the dropout/success factors at Medialogy bachelor programme and the risk group. Table 1 shows an overview of the data that we analysed for this project.

Table 1. Description of the data used in this project and where the data originates.

Data	Description	Data sources
General student data	Student demographics including gender, residence before studying and so forth.	STADS
High school grades	Student exit grades from high school.	STADS
AAU grades	Student grades at AAU, including re-exams.	STADS
Dropout status	Student enrollment status is either active or dropped out.	STADS and study secretaries
Study verification test (studiestartsprøve, SSP)	First semester student results from a questionnaire querying known dropout and retention factors from the literature. Medialogy BSc cohort 2017 only.	Moodle
Course performance	Online student behaviour in Virtual Learning Environments. Medialogy BSc cohort 2017 only.	Moodle, KhanAcademy, Peergrade

We got most of the data from STADS via QlikView or data handlers. Not all data is available on QuickView, and we had to contact data handlers who queried this from STADS database. The data from STADS is updated whenever a new data is available, whereas the QlikView data seems to updated each semester. Although we have not a full understanding of the STADS data entry system for grades and enrollment status, the data reveals some inconsistencies over the years, e.g. with course names and which date a grade is entered in STADS ('bedoemmelsesdato'). Some data is entered manually, e.g. by the study secretaries, who can follow specific STADS data entry manuals. The study secretaries at the first student year updates the dropout data in the end of each month and upload this to AAU's website[1] . "Q999" is a label of the students who are not active in a group on the semester but not yet dropped out from AAU on the first bachelor year. The study secretaries updates a list of active and Q999 students at the end of every month internally, but publish this online less frequently[2].

To collect demographic student data, we conducted a questionnaire for the SSP, including 15 categories consisting of 111 questions in total. The SSP questions are based on verified instruments and factors from various retention studies.

Lastly, we designed online activities in the introductory programming course, such as Moodle quizzes, which we downloaded manually. Merging the datasets from the different data sources could provide a better understanding of the risk group and success factors, but in this project we kept them separate.

[1] http://www.tnb.aau.dk/ansatte/Statistikker/

[2] http://fotoalbum.tnb.aau.dk/public/faces/fotoalbum/faggrupper.xhtml;jsessionid=9C04397FD4E961C3F81F63F8AD0F14B6

Results of the dropout and retention analysis

Predicting dropouts at Medialogy BSc (cohort 2012-2014).

- The current traffic light system[3] from Qlikview/STADS used by the study board to target failing students (i.e. students missing more than 5 ECTS points) does not work on the first two semesters. After the 1st semester every student missing any ECTS is at a very high risk of dropping out (risk is 74%). The 5 ECTS traffic light model works better after the third semester at Medialogy BSc.
- Medialogy students on the study plan from 2014 dropped out much earlier compared to students from the previous study plan, while the dropout rate by the end of an education was the same (see Figure 1).

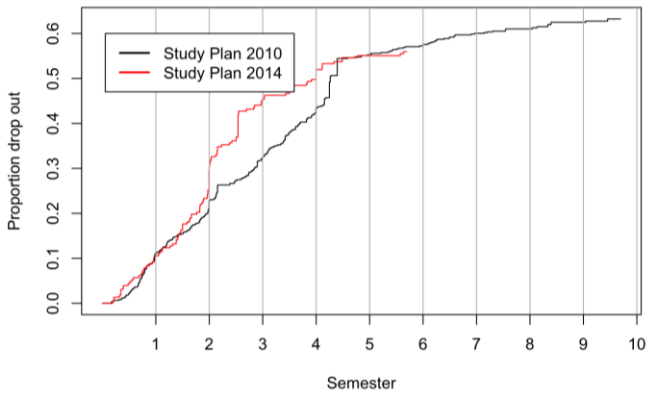

Figure 1. Subdistributions of dropout when stratifying according to the study plan.

Predicting first-year dropouts based on high school data and the SSP questionnaire (cohort 2017).

- AAU course grades (especially technical) on the first semester had a bigger influence on dropout risk compared to project grades in analyses of the cohorts starting from 2012 to 2014. Data analysed of student demographics and grades found factors related to dropout: qualifying education, GPA, level of math, and grade in math A, all from qualifying education.
- Dropout predictors of the 16.5% dropouts after the first semester (Feb. 2018) found from the SSP included: perceived academic abilities and studying/working hours. dropouts from cohort 2017. After the 2nd semester (June 2018) only perceived academic ability remained.

Predicting performance in the introductory programming course on the first semester (cohort 2017).

- Predictors of the introductory programming exam grades included: midterm exam scores, Khan Academy activity, and from the SSP high school trust, self-control, and personal trait comparison. The initial results using stepwise linear regression and including midterm exam scores: Khan Academy activity, self-reported programming experience, and a subset of self-reported dropout/retention factors [3]. When excluding the midterm exam: average self-assessment quizzes score, Peergrade submission score, self-reported high school trust, self-reported reasons for going to university, and self-reported self-control.
- Repeating the analysis with a lasso regression found a new combination of predictors: Peergrade combined score, self-reported high school trust, self-reported belonging uncertainty, self-reported personal trait comparison, and self-reported understanding of Medialogy. The different results in [3] and [4] suggests that the results are difficult to replicate with new data, e.g. for cohort 2018.

[3] https://www.kvalitetssikring.aau.dk/noegletal-kvalitet/procedure-vejledning-frafaldstruede-studerende

Identifying risk group on the first semester (cohort 2017)

- Cluster analysis based on the SSP in 2017 identified two student clusters. Main differences between them were in: attention to education, belonging uncertainty, education choice factors, financial support, grit, high school behaviour, personal trait comparison, and self-control.
- We identified risk group of 20 students based on midterm exam score and SSP score (December 2017). Interviews with this group identified complications, e.g. long commute times, bad study habits, and low self-esteem. By May 2018, nine of the 'risk group' students had dropped out, and three students were not part of a study group; whereas for the students considered 'out-of-risk' we had 23 dropouts and 15 students not in a study group.
- Based on a midterm exam scores in the introductory programming course, we identified 43 students requiring help and invited them for additional voluntary tutoring sessions. Uptake was a dismal 16%. Out of the 33 who later failed the course 30 had received an invitation (91% recall). Attending tutorials increased passing rates to 50%.

Similar to the findings in [2], our results show various reasons to drop out on the first year of Medialogy. Thus, using a variety of data to analyse student dropouts is necessary to improve the prediction quality. Only the students' perceived academic ability was a significant dropout predictor in cohort 2017, although we tested predictors from other scientific work.

Preparing and creating a web application prototype with dropout predictions

We created a web application prototype that shows a Medialogy BSc student's risk of dropping out in each semester. In this section, we describe the different steps in creating the prototype. Figure 3 shows an overview of the development steps, from data import to visualizing the dropout risks on the web application. We wrote the data import, data processing and prediction analysis in R statistics program and the web application in Python using Django and DataTables.

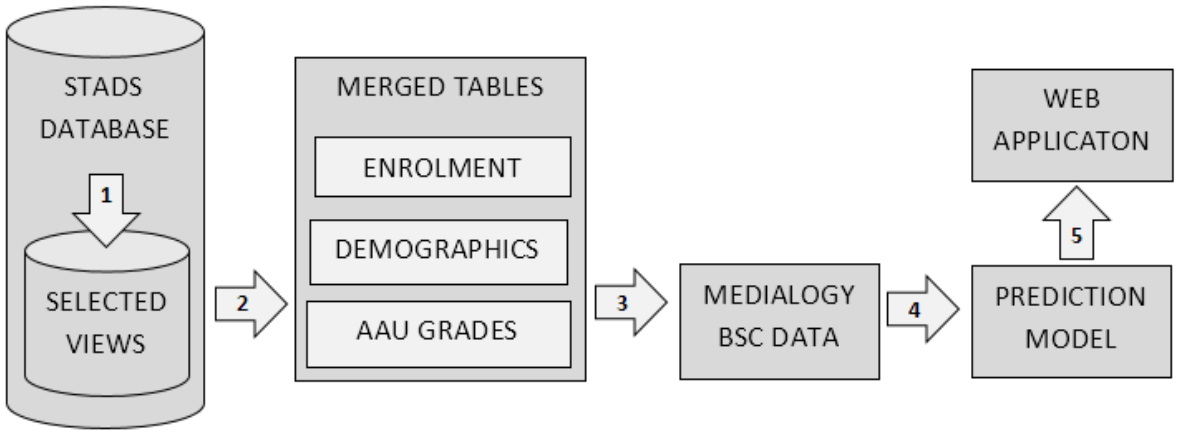

Figure 3. Development steps of creating the web application prototype. The labelled arrows describe: (1) Got access to selected views from STADS database. (2) Merged the selected STADS views for all AAU students. (3) Subsetted the merged tables to Medialogy BSc and merged it with Medialogy specific data, e.g. study plan data, SSP, and course performance. (4) Prediction model using the most recent and finished cohort as training data, i.e. cohort 2014. (5) Show the prediction results on the web application.

In addition, we gained experience in how the prototype can be expanded for more educations, and thus, be of value for more study boards by automating the process.

Planning data collection based on significant predictors

Prediction model used for the prototype: We fitted logistic regression models to data for the Mediology cohort starting in 2014, and the data was gathered after their sixth semester. SSP data and course performance data were not collected for this cohort. This data contains 295 students and 164 of these dropped out. The data also contains semester-wise information about how many ECTS the students have attained and their grade average. The prediction models are fitted only for students who started on the relevant semester (e.g. the information from the second semester can only be used for students who finished their second semester or higher). When only considering students who finished their fourth semester, too few students dropped out make a reliable model, hence we have four prediction models based on cohort 2014. All models were chosen from step-wise selection and shown in Table 2.

Table 2. Logistic regression models for predicting dropout trained on the Medialogy BSc cohort 2014.

Model information	Significant predictors
Prior enrolment	Exit math grade from high school.
After 1st semester	Number of attained ECTS points in technical courses on the first semester.
	Project grade on the first semester.
	Grade average of non-technical courses on the first semester.
After 2nd. semester	Number of attained ECTS points in technical courses on the first two semesters.
	Project grade on the first two semesters.
After 3rd semester	Number of attained ECTS points in technical courses on the first three semesters.
	Grade average of non-technical courses on the first three semesters.

Work on data import and access to STADS views

Data access inquiry of STADS views: enrolment data, enrolment/dropout status, student demographics, high school grades, and AAU grades.

Based on our project experiences, we recommend that other projects working with the STADS database collaborate with the relevant staff members from the beginning. We made efforts in locating the right people, communicating the project needs, and understanding what is possible with the STADS database. This was a huge effort because there was little institutional knowledge accessible to us as to who keeps what data in what formats and whether which of the data could be obtained and merged together. The STADS team would be able to provide more insights into the inner workings of what made it difficult for them to respond to our requests.

The staff members at AAU part of the process of getting access and importing views from the STADS database and Moodle were:

- AAU's data protection officer - Teia Melvej Stennevad: defined the rules for data usage in the project.
- Pernille Refstrup - field manager of Study Systems: approved our data access inquiry for the duration of the project and managed the discussion with the DPO.
- Gert Espersen - Team Lead, ITS: helped in the beginning stages of the project when we sought access to STADS.
- Anders Møller - database administrator: gave us access to the specific STADS views when going through Pernille as the gatekeeper and provided help in configuring the ODBC driver.
- ITS: Jens Sandberg Andersen - network officer: Assigned us fixed IP address for each machine and configured the firewall so we could connect to the STADS database and query its views from inside the AAU network .

- Birthe Riis Kennedy - daily support of the administration's running applications as STADS: supported us in getting an overview of the STADS views.
- Tina Funck (ITS Moodle) - provided help in setting up tables to temporarily store Moodle activity data for analytical purposes and tried out the learning analytics INSPIRE plugin, which turned out to be insufficiently configurable for AAU's purposes. The current version had to always run on the whole AAU Moodle database which was too resource consuming and smaller runs did not yield any usable prediction data.

Data processing: merged, subsetted, and created aggregations

We merged the STADS views for all AAU students to three different tables:

- Enrolment and dropout data for each student enrolment at AAU,
- Demographics and high school grades for each student, and
- AAU exam grades for each student activity.

We then subsetted the merged views into Medialogy BSc data to base the prediction model within the education. As mentioned in the previous section, we did not include course performance data (e.g. from Moodle) and the SSP responses, as the information was only collected for cohort 2017 and not for the training data.

The next step required creating aggregations of the AAU grades using information of the study plan versions and their study activities. We grouped Medialogy BSc grades into activity types per semester (i.e. technical courses, non-technical courses, elective courses, and semester projects) and made average grade aggregations. We created attempted and attained ECTS per semester.

As a final step, we performed a semester-wise prediction analysis for the Medialogy BSc student using a list of possible predictors and trained on cohort 2014. We then applied the prediction coefficients on cohorts younger than 2014 to compute the risk for each student.

Web application prototype

The web application is password-protected. The idea is to grant the study board members or other administrators access to their respective educations. Due to the general data protection regulations (GDPR), the prototype will be dormant working in March 2019.

We developed a user interface using a fake dataset for simulating the students at risk, see Figure 4. The user can select prediction models of different semesters, cohorts, and educations from a drop-down list, see Figure 5. The name of the predictors change for each model, and the user can hover over the predictor name to view more information, see Figure 6.

You can watch a short demo of the web application here: https://youtu.be/EnrIxPltxL8 and the web application code is available at: https://github.com/bastianilso/studentretention.

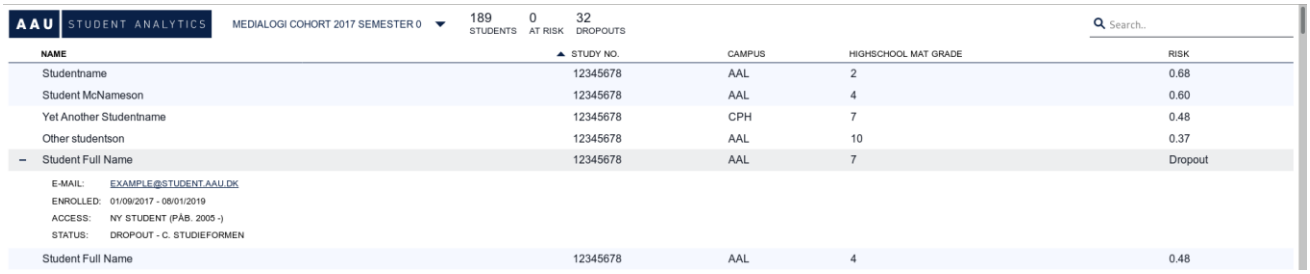

Figure 4. Web application interface of a student's dropout risk point for the selected education, cohort year and semester. Here are the prediction result for Medialogy BSc cohort 2017 using data prior their AAU enrolment (i.e. demographics and high school grades).

Figure 5. Drop-down list of the possible prediction results, grouped by education, cohort, and semester.

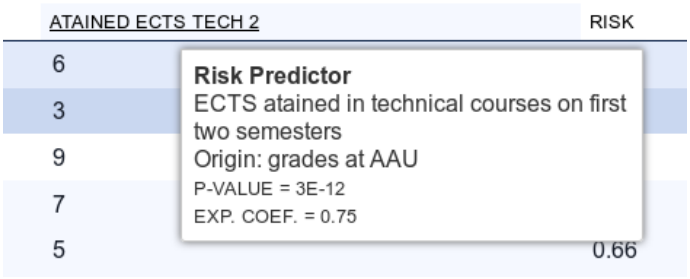

Figure 6. Hover-over interaction of the predictor name (here attained ECTS TECH 2) shows a predictor description, data source, and prediction model results.

Limitations

We trained the prediction model on the last cohort finishing their bachelor following the same study plan version, which avoids the problem of active students still being able to drop out. However, with the year gap, younger cohorts might have new circumstances that are not currently reflected in the model, e.g. data logging of online course activities/performance and the SSP.

We excluded SSP data and course performance data in the prediction model training, as the data collection of this information started in 2017, and we do not have enough dropout students to create reliable models.

The course ID changes for each year on STADS database, and the corresponding string name is entered manually, e.g. by study secretaries. We found no ID tracking the course names across years, and without it, there is little chance in locating the old course name based on string comparison alone. Hence, it would require some manual work to keep consistency over the years with the string matches, so the system can use the predictors on different cohorts.

The prototype currently works only for the Medialogy BSc education. Our exam activity aggregations of the grades only work for Medialogy BSc, as we used the study plan to group these in technical and non-technical courses. This also required the above mapping of various different spelling of courses in STADS. Other educations might find different relevant SSP questions or group grades into course grades and project grades per semester. These different predictors would need to be included into the database prior to modeling the data sets.

The problem with the course IDs as described above is reflected when using the exam activities as predictors. Although, the total attained ECTS points can easily be tracked for each student, the tracking of ECTS points of a particular exam activity becomes more complicated. In addition, the merit system might also have some challenges that we have not explored in this project.

Dissemination and embedding

We have presented our findings and designed tools (automated SSP and course progress feedback) to the learning analytics ‘frafald’ group headed by Nils Peter Uhre at AAU who’s reporting to Det Strategiske Uddannelsesråd. We are involved in the re-design the individual tutorials in introductory programming course.

We published and presented our findings from the programming course at the 2018 conference for smart learning ecosystems and regional development (SLERD). The 2017 cohort was very interested in the phenomenon of dropout, which we presented at P0 and P1 start-ups. We have also shared our insights with the study board and other relevant stakeholders. We would be happy to demonstrate or explain how the tool would be used to interested parties.

While we did not evaluate it, we believe that one of the biggest unintentional contributions of the project might be that raised the awareness of the major stakeholders involved in the topic of learning analytics about the challenges in the current infrastructure and processes in place at AAU.

Conclusion

The project addressed the themes of PBL, digitization, and student retention during the first year of study. It aims at enabling relevant stakeholders such as semester coordinator, student counsellors, study boards, and putting retention on better footing by predicting failing students as early as possible to allocate resources to them. The prediction analysis presented in this report provide initial findings in support of additional strategies for targeting struggling Medialogy BSc students. While the prediction model can identify students at risk, we could not replicate the same combination of predictors in the model when using different prediction approaches, suggesting that we should use the predicted risk as guidance only. Our results is similar to other studies in the retention literature which has low agreement between the studies [5, 6, 7], between various circumstances across educations, faculties, universities, countries, and cultures.

For continuing this work, we suggest that dropout prediction models should at least be created education-wise, but considering our prediction results, the models could also be created semester-wise. We also suggest continuing logging the general student data, high school grades, AAU grades, dropout status, students' self-reports from the SSP, and course performance (e.g. from online activities). In particular for the SSP and the course performance data, we need data collection for more than one cohort/year and more than one year within a cohort. Automating every step of the prediction analysis and collecting data during the semester (e.g. online course performance), could be useful for identifying struggling students at any point in the semester, and it is something that we only have touched upon in this project.

Future work

In the future, we hope to implement significant dropout predictors for identifying struggling students and incorporate additional sources of relevant information, such as Moodle course activity. Future analysis could focus more on analysing the interaction effects of the collected variables.

Providing individualized counselling based on data: We experienced low usage of the student counsellors at the Medialogy education, even though many students failed the technical courses and should have contacted a counsellor. We therefore propose that students can choose to get individualized counselling based on data. It can be opt-in for the student, choosing which data sources to include in their counselling (i.e. from STADS, Moodle, or SSP). In a conversation with the students struggling in the introductory programming course, we found that first year students would prefer to have the study counselling earlier in a semester rather than in the end. Such counselling is also offered at other universities, e.g. the University of Eindhoven University, at the first year of the education, to help students choosing the right education.

AAU will need to make efforts towards understanding to what degree and how students need to consent to this form of assistance through learning analytics, e.g. to abide by GDPR regulations.

References

[1] Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*.

[2] Bøgelund, P., Justesen, K., Kolmos, A., & Bylov, S. M. (2016). *Undersøgelse af frafald og fastholdelse ved medialogi og andre uddannelser ved det Teknisk-Naturvidenskabelige Fakultet 2015-2016: Arbejdsrapport Nr. 1*. Aalborg Universitet.

[3] Christensen, B. C., Bemman, B., Knoche, H., & Gade, R. (2018, May). Identifying Students Struggling in Courses by Analyzing Exam Grades, Self-reported Measures and Study Activities. In *Conference on Smart Learning Ecosystems and Regional Development* (pp. 167-176). Springer, Cham.

[4] Christensen, B. C., Bemman, B., Knoche, H., & Gade, R. (2019). Pass or Fail? Prediction of Students' Exam Outcomes from Self-reported Measures and Study Activities. Paper submitted for publication.

[5] Herzog, S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in higher education*, 46(8), 883-928.

[6] Lassibille, G., & Navarro Gómez, L. (2008). Why do higher education students drop out? Evidence from Spain. *Education Economics*, 16(1), 89-105.

[7] Touron, J. (1983). The determination of factors related to academic achievement in the university: implications for the selection and counselling of students. *Higher Education*, 12(4), 399-410.