Aalto University
School of Business

CHALLENGES OF IMPLEMENTING LEARNING ANALYTICS IN THE CONTEXT OF BLENDED UNIVERSITY COURSES

A reflexive case study of a course completion prediction modelling project

Master's Thesis
Sini Luostarinen
Aalto University School of Business
Marketing
Spring 2021

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Sini Luostarinen

Title of thesis Challenges Of Implementing Learning Analytics In The Context Of Blended University Courses: A Reflexive Case Study Of A Course Completion Prediction Modelling Project

Degree Master of Science in Economics and Business Administration

Degree programme Marketing

Thesis advisor(s) Alexei Gloukhovtsev

Year of approval 2021 **Number of pages** 42 **Language** English

Abstract

As during the last decade, the use of learning management systems (LMS's) as course platforms became the norm in courses of higher education, and vast quantities of learning behavior data was thus collected, also the practices of educational data mining and learning analytics began growing significant popularity in the science community. For instance, much research on predicting course outcomes, such as course dropouts and course grades were conducted, in both; the online and blended learning contexts. However, research on the challenges which a learning analyst faces while implementing a learning analytics project in a higher education institution, were nearly nonexistent.

When working with higher education institutions which haven't yet developed many internal processes for supporting the implementation of learning analytics, learning analysts may face challenges with data access permission and data quality, for instance. Thereby, their projects may suffer from schedule delays or deteriorated results. With a reflexive case study conducted for Aalto University in Finland, this research identifies the main challenges learning analysts face while implementing each stage of the project life cycle. The case project used in this reflexive study is a course completion prediction modelling project that utilizes learning data of marketing students.

The challenges identified during the case project occurred only in the data collection and pre-processing stages. First of all, data privacy issues initiated a legal process which due to a heavy work load, slightly delayed the schedule of acquiring permission to accessing student data. Secondly, hand-picking data manually from two databases and integrating that data from two data systems lead to a missing data problem. The data quality also proved to be rather poor, due to unstandardized and partially nonexistent data collection in the LMS. Finally, most of the modelling data had to be eliminated due to incomparability and inoperability challenges, and the prediction model was conducted successfully only for one marketing course.

Based on the experienced challenges, this research introduces a set of propositions for managers of higher education institutions, which help overcome these challenges by improving the institution's internal processes and developing new ones. The propositions are: 1) Create a standardized process for applying for data access; 2) Adopt a learning analytics information system to ease data integration; 3) Harmonize course and course-site structures to enable automated variable construction; and 4) Organize courses in a way which supports data generation. Considering these improvement propositions, this research can help higher education institution managers both improve the life cycle times of learning analytics projects, and the quality of the project results, while simultaneously enhancing the learning of their students.

Keywords learning analytics, educational data mining, project implementation challenges, course completion prediction modelling, higher education

Aalto-yliopisto
Kauppakorkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Maisterintutkinnon tutkielman tiivistelmä

Tekijä Sini Luostarinen

Työn nimi Oppimisanalytiikan implementoinnin haasteet sulautuvassa oppimisympäristössä: Refleksiivinen tapaustutkimus kurssin suorittamisen ennustemallinnusprojektista

Tutkinto Kauppatieteiden maisteri

Koulutusohjelma Markkinointi

Työn ohjaaja(t) Alexei Gloukhovtsev

Hyväksymisvuosi 2021 **Sivumäärä** 42 **Kieli** Englanti

Tiivistelmä

Kun viime vuosikymmenen aikana oppimisen hallintajärjestelmien käytöstä kurssialustoina tuli korkeakoulutuksen standardikäytäntö ja suuria määriä oppimiskäyttäytymisdataa alettiin kerätä entistä enemmän, oppimisanalytiikan ja opetuksellisen tiedonlouhinnan suosio alkoi merkittävästi kasvaa tiedeyhteisöissä. Oppimisanalytiikan tutkimusta tehtiin mm. kurssien lopputulemien ennustamiseen, kuten kurssin keskeyttämiseen sekä kurssiarvosanoihin liittyen, niin online- kuin sulautuvan oppimisen ympäristöissä. Kuitenkin tutkimus siitä, millaisiin haasteisiin oppimisanalyytikot törmäävät, kun he toteuttavat oppimisanalytiikan projekteja korkeakouluinstituutioissa, on tänä päivänä lähes olematonta.

Kun oppimisanalyytikot työskentelevät korkeakoulujen kanssa, jotka eivät ole vielä valmistelleet sisäisiä prosessejaan oppimisanalytiikkaprojekteja tukeviksi, oppimisanalyytikot voivat törmätä haasteisiin mm. dataan käsiksi pääsyn tai epälaadukkaan datan kanssa. Näin ollen heidän projektinsa voivat kärsiä viivästymisistä tai epäkäytännöllisistä tuloksista. Tämä tutkimus keskittyy identifioimaan suurimmat haasteet mihin oppimisanalyytikot kussakin projektin vaiheessa törmäävät toteuttaessaan projektejaan. Tutkimus on tehty refleksiivisenä tapaustutkimuksena (reflexive case study) Aalto-yliopistolle ja tapaustutkimuksena on käytetty kurssin suorittamisen ennustemallinnusprojektia, jossa on hyödynnetty markkinoinnin opiskelijoiden oppimisdataa.

Haasteet, jotka tapaustutkimuksessa tunnistettiin, esiintyivät vain datan keräämis- sekä esikäsittelyvaiheissa. Projektin alussa opiskelijadatan tietosuojahaasteet ajoivat aloittamaan monimutkaisen juridisen luvanhakuprosessin, joka suuren työmääränsä takia viivästytti hieman luvan saamista dataan käsiksi pääsyyn. Seuraavaksi oli edessä datan manuaalinen poiminta sekä datan integrointi kahdesta opiskelijadatanjärjestelmästä yhdeksi datasetiksi, joka johti suureen määrään puuttuvaa dataa. Myös datan laatu paljastui heikoksi epästandardisoidun sekä paikoittain jopa olemattoman datankeruun takia oppimisen hallintajärjestelmässä. Lopulta suurin osa datasta piti karsia kokonaan epälaadukkaan datan sekä eri kurssien välisen vertailukelvottomuuden takia, ja ennustusmalli pystyttiin onnistuneesti rakentamaan vain yhdelle markkinoinnin kurssille.

Perustuen koettuihin haasteisiin, tämä tutkimus tarjoaa myös kehitysehdotuksia korkeakoulujen johtajille – niin sisäisten prosessien parantamiseksi, kuin uusien prosessien kehittämiseksi ja näin ollen haasteiden ylitsepääsemiseksi. Tutkimuksen kehitysehdotukset ovat: 1) Luo standardisoitu prosessi dataan käsiksi pääsyn luvan hakemiselle ja myöntämiselle; 2) Ota käyttöön oppimisanalytiikan tietojärjestelmä helpottamaan ja tehostamaan datan integrointia; 3) Harmonisoi kurssien ja kurssisivujen rakenteet niin, että muuttujien luominen voidaan toteuttaa automatisoidummin; 4) Järjestä kurssit niin, että edesautat datankeruuta. Näiden kehitysedotusten myötä tämä tutkimus auttaa korkeakoulujen johtajia niin tehostamaan oppimisanalytiikkaprojektien läpimenoaikoja, kuin parantamaan niiden laatua, ja näin ollen myös tehostamaan opiskelijoiden oppimista.

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Maisterintutkinnon tutkielman tiivistelmä

Avainsanat oppimisanalytiikka, opetuksellinen tiedonlouhinta, projektin implementoinnin haasteet, kurssin suorittamisen ennustemalli, korkeakoulu

Table of Contents

List of Tables

List of Charts

1. Introduction

It's widely known in higher education, that dropping out in the middle of a university course can impose difficulties for students' future studies. Dropped courses can pile up and lead to excessive quantities of work for future school periods, triggering a chain reaction that occurs one period after another, ultimately deteriorating students' wellbeing and study outcomes. In Finland, more pressure to graduate within a specific time frame is put on students due to a limit on the amount of student allowance months granted, which is set by the government. Thereby, keeping up with the recommended study schedule is oftentimes important to students from a financial and social standpoint, too. To decrease students' course dropouts, education institutions can utilize learning analytics. During the past decade, learning analyst researchers have conducted more and more studies utilizing large quantities of student data, and developed data mining approaches (e.g. Olivé et al., 2020; Wladis et al., 2014; Yang et al., 2013) and retention strategies (e.g. Burgos et al., 2018; Hawkins et al., 2013) to decrease course dropouts, thus offering education institutions opportunities to improve students' learning, study schedule stability and well-being.

Higher education institutions of today commonly arrange courses within learning management systems (LMS), or e-learning systems, which are online platforms for organizing courses, distributing study materials, organizing group discussions and receiving students' assignment submissions, for instance. From the viewpoint of learning analytics, one important upside of LMS' is that they collect large amounts of learning data that can be used in educational data mining (e.g. Daniel, 2015; Gitinabard et al., 2019; Zafra & Ventura, 2012). As during the past few years the use of LMS's has expanded rapidly (e.g. Cohen, 2017; Gitinabard et al., 2019; Zafra & Ventura, 2012), researchers have accessed much larger amounts of learning-related LMS-data, and have been able to use it for much more complex educational data mining purposes. As learning analytics refers to collecting educational data and analyzing it to predict study performance or identify risks, educational data mining is concerned with developing methods for exploring those large amounts of educational data (Romero et al., 2008). Learning

analytics can thereby provide useful insights for educational instructors and administrators which can be used to improve students' learning.

To decrease course dropouts, learning analysts can try to predict those students who are most likely to drop out of a course with prediction models, and course faculties then can target those students with retention interventions. Course completion prediction modelling is a practice of learning analytics, which aims to predict which students are at-risk of dropping out based on chosen student and course-related data. A successful course completion prediction model can help identify those students at-risk early on in the course, and therefore retention interventions such as initiating tutoring, tailored assignments or just adapting teaching to students' needs (Daniel, 2015), can be targeted at those students, to retain them on the course and thus to enhance their learning. The aim of course completion prediction modelling, just like any learning analytics, is the improvement of students' learning, which can be achieved by planning and executing changes on teacher-student interactions and teaching methods, based on learning analytics insights.

Although learning analytics has been used for many successful projects with practical implications to students' learning and teaching methods, learning analytics projects oftentimes involve challenges and difficulties, too. The implementation of learning analytics projects oftentimes involves very complex processes (Nguyen et al., 2020). For instance, Daniel (2015) argues, that especially when implementing data mining techniques for big data in higher education, one challenge is to get people to accept learning analytics as the new information source to base new process development on. Dyckhoff et al. (2012) and Krüger et al. (2010) on the other hand highlight the problem of poor data quality in LMS's. Another challenge is the lack of interoperability amongst different institutional data systems (Daniel, 2015; Nguyen et al., 2020). Moreover, data integration problems, and the possible consequent loss of data are issues which can become obstacles and deteriorate the results of learning analytics (Daniel, 2015). As researchers have shed light on the difficulties that may arise in the implementation of learning analytics, very little research focusing on those difficulties has been conducted in context of course completion prediction modelling thus far.

As multiple previous studies have examined course completion prediction modelling in the higher education context and resulted in a lot of useful information on the best prediction

practices, research on the challenges and difficulties that may occur in the data collection, data handling and modelling processes is still rather non-existent. Research on these challenges is important, since not recognizing them and not knowing how to avoid them might lead to a delayed project schedule, unsatisfactory results or even a failed modelling project. If educational institutions recognize these challenges and improve the implementation process of learning analytics, more efficient and impactful learning analytics could be conducted, and students' learning could be improved. Improved learning could show as graduating faster, enhanced course grades, and improved ability to take on multiple courses simultaneously. Thereby, the goal of this paper is to help university faculties to recognize and tackle these challenges; to reflect on the challenges in each stage of a course completion prediction modelling project, and to recommend solutions for those challenges.

The research questions of this study are:

1. What are the challenges in implementing a course completion prediction model of blended marketing courses in higher education of today?
2. What issues should university processes take into account concerning development, implementation and use of learning analytics?

To identify the challenges of learning analytics implementation, and moreover, to find the best solutions to overcoming these obstacles, I implement a case study of course completion prediction modelling. The study shows evidence that learning analytics projects can suffer from some obstacles in the implementation process – a comprehensive course completion prediction model is initially planned to be executed with data of all Aalto University's volitional marketing courses of 2018-2020, but eventually, is only executed for one course due to challenges with data integration and quality. To provide as useful insights as possible for universities, the learning context of this case study was the most common learning context in higher education; blended learning, indicating that the courses examined included both, e-learning and face-to-face in-class learning, which posed its own challenges on the modeling process as well.

In chapter 2. Literature review, I review the literature on previous course outcome prediction models in the higher education context and obstacles of learning analytics, such as the data-related challenges of the blended learning context found in literature thus far are being

addressed. In chapter 3. Methodology, I explain the approach, methods and the research gap of this reflexive case study. In chapter 4. Findings, I introduce the case project and its methodology. More importantly, I address the findings – challenges and obstacles that were found in each project stage, thoroughly in this part. In chapter 5. Discussion, I reflect on the findings of about implementation challenges, suggest solutions and discuss limitations. I conclude the paper with a summary and some useful recommendations to benefit learning analysts' work in chapter 6. Conclusions.

2. Literature review

The emerging phenomenon of learning taking place more and more in online environments such as LMS's today (Daniel, 2015), has fueled the practice of educational data mining during the past few years. Countless data mining studies using data of student behaviors, especially on LMS platforms exist today and give recommendations to learning analysts and administrators of higher education. One emergent branch of learning analytics is prediction modelling of student's course outcomes. However, during the process of implementing such prediction modelling projects in higher education institutions, some pitfalls and obstacles may emerge that may lead to a delayed project, imperfect results or even a failed project. This literature review will first look at previous course outcome prediction model studies in the higher education context and their findings, and after that move on to reviewing previous literature related to the challenges of implementing learning analytics in a higher education institution.

2.1. Course outcome prediction modelling

In learning analytics literature, the two approaches to predicting course outcomes are student performance prediction, which predicts the course grade or success of students; and course completion prediction, which predicts classification to passed and dropped students. Course completion prediction modelling is a method for identifying students at-risk of dropping out

based on their early behavior and other information. Accurate prediction in student performance prediction modelling enables education faculty to identify the students most likely to result in bad study performance and target supportive actions towards them early on. In course completion prediction on the other hand, the aim is to identify students at-risk of dropping out of the course, and targeting them with supportive actions to improve their learning and increase the likelihood of retention (Cohen, 2017; Olivé et al., 2020; Wladis et al., 2014).

Research on course outcome prediction today provides a countless number of unique models which can predict student dropout in different learning contexts, using different types of variables and different modelling techniques. Many research approaches aim to test if certain aspects of students' behavior can be used to predict course outcomes – for instance Lee (2018) found that uninterrupted time on-task, and Joksimovic & Gaševic (2015) that social presence in online discussions have predictive power on course outcomes. Course outcome prediction research has to a wide extent focused on identifying the variables with most predictive power, and thus finding ways to predict course outcomes as accurately as possible. However, student behavior variables aren't the only variables that have been researched in course outcome prediction studies – also course-level variables (Wladis et al., 2014, 2017) and student demographics (Ibrahim & Rusli, 2007) have been found to have significant predictive power.

Moreover, a lot of research on different types of modelling techniques and approaches have been conducted. As many established techniques exist in course outcome prediction modelling, logistic regression probably being the most commonly used one (used e.g. by Hawkins et al., 2013; Jiang et al., 2014; Lee, 2018; Taylor et al., 2014; Wladis et al., 2014), many researchers have developed completely new approaches to identifying churners early on during the course. For instance Halawa et al. (2014) were able to create their own red-flagging system which used active and absent mode predictors, and Cohen (2017) a dropout alert index, to predict student dropouts. Also comparative research on different modelling techniques has been conducted, to find out the best modelling techniques or examine the advantages and shortfalls of each technique. For instance, Ibrahim & Rusli (2007) compared artificial neural network, decision tree and linear regression techniques for predicting student performance, and found out that artificial neural network outperformed the others.

Course outcome prediction models can be divided in at least three categories by the learning environment context; online, offline, and blended learning. The vast majority of course completion prediction models thus far has occurred in studies of the online learning context (see Hawkins et al., 2013; Wladis et al., 2014, 2017), and especially in MOOC (Massive Open Online Course) contexts (see Halawa et al., 2014; Y. Lee, 2018; Olivé et al., 2020; Taylor et al., 2014; Yang et al., 2013). Course completion prediction studies in the blended learning context on the other hand, exist quite limitedly. The share of student performance studies on the blended context on the other hand, exist more widely than in the course completion prediction context (see Gitinabard et al., 2019; Ibrahim & Rusli, 2007; Zafra & Ventura, 2012). Online course context studies such as MOOC studies seems also to be quite large in the student performance prediction area (see Jiang et al., 2014; Macfadyen & Dawson, 2010).

To create a prediction model with high prediction accuracy, including variables in it which are known to have high predictive power, is crucial. As many great findings concerning predictive variables have already been made in previous studies, the following sections will review these studies and the variables that they have found to have best predictive power. Moreover, the predictive accuracy of those models and the techniques used to achieve the results are discussed. The section 2.1.1. focuses on the findings of course outcome prediction studies in the online and MOOC context, whereas 2.1.2 addresses the context of blended courses. Both type of prediction studies are addressed; student performance prediction and course completion prediction. A summary of course outcome prediction modelling studies in previous literature, their modelling techniques and their findings of best outcome predictors, are listed in Table 1.

2.1.1. Outcome prediction in online courses

Online courses and MOOCs especially, which use LMS's which collect great quantities of rich student behavior data, have been researched in literature quite extensively thus far. MOOCs are online courses that are open for everyone and oftentimes have a massive participant count. Research on MOOCs has proven that students' online behaviors can be used to predict their performance or probability to complete a course with high accuracy (see Ibrahim & Rusli, 2007; Jiang et al., 2014; Y. Lee, 2018; Olivé et al., 2020; Yang et al., 2013; Zafra & Ventura, 2012).

However, findings of course outcome prediction studies in the MOOC context cannot be directly applied to the university context. MOOC completion rates are relatively low even among students who intend to complete the course, which could stem for instance different motivation levels and backgrounds of MOOC students and postsecondary level students (Jordan, 2015). This implies that there might be quite different factors affecting course outcomes in the MOOC context than in the postsecondary course context. The learning settings are also very different from each other. As the norm of postsecondary learning today is the blended course, student-student and student-teacher interaction is much different between MOOCs and postsecondary learning – in the blended course context, students get to interact with teachers and peers face-to-face.

Many online course and MOOC researchers have examined course outcome prediction based on individual-level data; students' demographics data and their study behavior data. Student demographic variables used in course outcome prediction are for instance the age and education level of students. Study behavior data used in such prediction models on the other hand, are collected in the LMS environment and could include the number of assignments submitted, number of forum discussion posts, length of forum posts, ratings on forum posts, videos skipped, time spent online, time spent absent, uninterrupted time on-task, number of quizzes attempted and the timestamp of when the student joined the course (see Halawa et al., 2014; Jiang et al., 2014; Y. Lee, 2018; Macfadyen & Dawson, 2010; Olivé et al., 2020; Yang et al., 2013).

Olivé et al. (2020) found the following LMS features important in their MOOC completion prediction study; positive ratings in database entries, positive ratings in forum posts, the student attempted at least one quiz and percentage of quizzes attempted in the course. They reached an accuracy of 88.81% with their neural network model trained on past course data. Halawa et al. (2014) on the other hand used LMS variables such as video skip, assignment skip, quiz grade, length of active mode and length of absent mode in their MOOC completion prediction model. The absent mode predictor, which flagged students who had been absent over 2 weeks, was found to be the most predictive one as it was able to pick up over 90% of dropouts in most of the courses.

Jiang et al. (2014) however did not study course completion prediction but student performance prediction and used two logistic regression models based on study behavior and student

characteristics data from the 1st week of the MOOC. They used social network degree which measures the level of "centrality" or social integration, the number of completed assignments, the average quiz score and education background as predictors and reached a model accuracy of 92.6% and 79.6% for their two models. Assignment submission and quiz completion seem to be very important factors in MOOC outcome prediction. In addition to Jiang et al. (2014), also Taylor et al. (2014), Macfadyen & Dawson (2010) and Olivé et al. (2020) have found such results. In fact, the MOOC completion prediction model by Taylor et al. (2014) implies that student problem submission engagement variables have the most predictive power on dropouts.

A finding occurring in many online course outcome prediction studies is the importance of student interaction and social network behavior as predictors. Macfadyen & Dawson (2010) for instance examined student success prediction modelling with LMS variables in the context of a small online course, where they classified students in to two groups based on their risk of course failure, and found that the number of discussion messages posted, and number of mail messages sent, were among the most important factors in their logistic regression model. These findings point towards the importance of interaction variables. They were able to reach a total accuracy of 73.7%. Yang et al. (2013) utilized similar study activity predictors in their MOOC study of course completion prediction, and also found that the among the most important features in the model were interaction-related variables such as discussion forum post length and how much they replied to peers in discussion forums.

Jiang et al. (2014) found that "centrality", the number of connections that each student has with other students can be used to predict student performance. Kovanovic et al. (2014) support this finding with their findings of students' social capital, and found that the type of students' interactive social presence (e.g. affective, interactive, or cohesive) was highly correlated with their social capital, which on the other hand affects their course performance. In an online course context study, Joksimovic & Gaševic (2015) found that course grades can be predicted with variables describing social interactions, such as continuing a thread, complimenting other users, and expressions of appreciation. Hawkins et al. (2013) in an online high school study however, found that the quality and frequency of interaction had a significant impact on course completion but not on the grade awarded. The importance of repeated and frequent social interactions were

also acknowledged to have a decreasing impact on course dropout rates in a MOOC context (Sunar et al., 2017).

2.1.2. Outcome prediction in blended courses

LMS variables are used heavily in prediction models of the blended course context, too. A study by Cohen (2017) – one of the very few course completion prediction studies in the blended learning context today, found that students most likely to dropout became somewhat absent from the course LMS before dropping out. More importantly, they were able to find the dropout's early traces in the LMS log data, which they used for modelling. On the other hand, the true negatives prediction accuracy from the three courses tested was only 66%, which is less than many of the online and especially MOOC studies have reached (see Macfadyen & Dawson, 2010; Olivé et al., 2020; Yang et al., 2013), and indicates that the absence of offline-learning data from these blended courses might play a part in the low accuracy. However, in their study, variables like the number of student actions, average of online actions in relation to the average of other student's actions, and average of all student activity days, were found to be predictive.

Gitinabard et al. (2019) examined three student performance prediction models in the blended course context; Random Forests, Support Vector Machines (SVM), and a Logistic Regression; which categorized undergraduate students in to two performance groups based on their study habits and social relationships: A- or above and B+ or below. The most important features in their model were amount of time spent on LMSs, number of online actions performed, sessions generated, focusing on one tool at a time per session, and number of questions answered in the discussion forum. Zafra & Ventura (2012) in their blended course study, used LMS data related to completing assignments, participating in discussion forums and taking quizzes, to predict the students' final performance, which was measured by classifying students to categories of passed and failed. The LMS variables used in these blended course studies are very similar to the ones used in the online and MOOC contexts, which implies that the same types of factors, such as assignment submission and discussion forum activity can be used in predicting dropouts also in the blended course context.

Ibrahim & Rusli (2007) on the other hand had a very different approach; they predicted the Cumulative Grade Point Average (CGPA) of undergraduate students in a blended course context, by not using LMS-data, but using general student information as independent variables; information technology application knowledge, previous school type, general programming knowledge, and family financial status. They tried three different models; artificial neural network, decision tree and linear regression, which all produced over 80% accuracy. This finding indicates the importance of general student information to course outcome prediction models.

Study	Model type	Context	Technique	Predictors of dropout
Wladis et al. (2014)	Course completion prediction	Online and face-to-face community college courses	Binary logistic regression	Student's attendance motivation (elective/ distributional / major requirement), course medium, course difficulty level, course type
Wladis et al. (2017)	Course completion prediction	Online and face-to-face community college courses	Multi-level logistic regression	Student's attendance motivation (elective/ distributional / major requirement), course medium, course difficulty level, course type
Halawa et al. (2014)	Course completion prediction	MOOC	A red-flagging system with active and absent mode predictors	Skipping videos, skipped assignments, lag from the course rythm, assignment performance
Olivé et al. (2020)	Course completion prediction	MOOC	Neural network	Positive ratings in database entries, positive ratings in forum posts, the student attempted at least one quiz, percentage of quizzes attempted in the course
Y. Lee (2018)	Course completion prediction	MOOC	Binary logistic regression	Frequency and the duration of uninterrupted time-on-task

Taylor et al. (2014)	Course completion prediction	MOOC	Binary logistic regression	Features involving student problem submission engagement & features involving inter-student collaboration
Yang et al. (2013)	Course completion prediction	MOOC	Survival analysis	Beginning the MOOC during 1st week, discussion post duration & authority (engaging other students in discussions)
Cohen (2017)	Course completion prediction	Blended university courses	Dropout alert index & Mann–Whitney analyses	Number of all student actions, relative average of actions in relation to others & average of activity days
Hawkins et al. (2013)	Course completion & performance prediction	Online high school courses	Hierarchical logistic regression	Interaction quality, interaction frequency, feedback, procedural- and social interaction
Jiang et al. (2014)	Student performance prediction	MOOC	Binary logistic regression	E.g. assignment performance & the degree of social integration during 1st week
Macfadyen & Dawson (2010)	Student performance prediction	Online university courses	Binary logistic regression	E.g. number of discussion messages posted, number of mail messages sent, and number of assessments completed
Gitinabard et al. (2019)	Student performance prediction	Blended university courses	Random Forests, Support Vector Machines, Logistic Regression	Time spent & activity in browser and study sessions, the number of sessions, the number of homogeneous sessions, social activity on Piazza
Ibrahim & Rusli (2007)	Student performance prediction	Blended university courses	Artificial Neural Network, Decision Tree, Linear Regression	Students' demographic profile & the cumulative grade point average (CGPA) for the first semester

Zafra & Ventura (2012)	Student performance prediction	Blended university courses	Multi-instance grammar guided genetic programming, G3P-MI	E.g. time spent in quizzes, assignments and forum section & number of quizzes, assignments submitted, and forum posts posts sent and read

Table 1. Predictors of course completion from previous studies

2.2. Challenges of implementing learning analytics

As learning analytics researchers have studied course outcome prediction widely from various perspectives, they have quite extensively failed to examine the challenges of implementing course outcome prediction studies, or studies of learning analytics overall. The studies that do exist and present some challenges related to implementing learning analytics, are mostly limited to examining the challenges through the viewpoint of data usage, more specifically big data usage in higher education analytics (Daniel, 2015) and LMS-data usage in educational data mining (Romero et al., 2008). Consequently, the challenges that have been recognized in literature thus far, are not presented comprehensively throughout the whole project life cycle. Moreover, studies focusing on the challenges of implementing learning analytics in the blended context don't exist at all and pose a large research gap. The following chapter 2.2.1. will discuss the findings of existing previous literature on the challenges of the blended courses in learning analytics, and 2.2.2. will then examine previous literature on the challenges that educational data systems, such as LMS' pose on the implementation of learning analytics.

2.2.1. Challenges of the blended learning context in learning analytics implementation

The reason for the high volume of online and MOOC outcome prediction studies may lie in the availability of data and the ability to measure students' behavior. According to Gitinabard et al. (2019) blended courses, which don't provide as rich and comprehensive data sets as MOOCs, haven't been researched nearly as much, partly because of the challenges the blended context

sets. For instance, LMSs of blended courses cannot capture students' interactions in classrooms or group work, and thereby this individual-level behavior data cannot be included in the prediction models. As many researchers use LMS-variables such as time spent on-site, and amount of assignment submissions in their student outcome prediction models (see Gitinabard et al., 2019; Macfadyen & Dawson, 2010; Zafra & Ventura, 2012), they are missing out on variables such as time spent on face-to-face lectures, score of participation on class, and different student-student interaction features in group work sessions. As informative data is not being collected from each phase of the learning process, the final prediction model is more likely to become less accurate and thereby not as useful as preferred.

Gitinabard et al. (2019) also highlight, that in order to build a useful prediction model, we must be able to train the model on one blended learning course offering, and then use it on other courses. Thus, the model has to rely on variables which occur the same way in each course the model is used upon. One problem with prediction modelling in blended learning is, that the requirements of student participation in face-to-face activities and online activities may vary significantly between courses. As oftentimes students' offline behavior data is not continuously collected in learning institutions, the courses which involve a heavier load of offline activities and less online activities, puts learning analysts in a situation where there's not enough data to build a course outcome prediction model. Moreover, if a model is trained on one blended course offering with a large percentage of measurable online activities, and then applied on a course with a low percentage of measurable online activities, the prediction accuracy of the model may suffer remarkably due to the lack of data. Thereby, as offline activity cannot be measured at least with the same accuracy as online activity, a prediction model, usually trained on online-behavior data and student information data, may lead to poor results if applied on other courses.

2.2.2. Challenges with educational data systems in learning analytics implementation

The use of online tools in postsecondary education has expanded rapidly in previous years, which has lead to an extensive shift from traditional university courses to blended courses and online courses. Blended courses with their mixed use of online learning technologies and face-to-face learning, are the typical university courses of today. Universities use online learning

management systems (LMSs) such as Moodle, to distribute learning materials, give grades, enable group discussions and to receive assignment submissions. The advantage of these learning management systems is not only that they support students' learning, but also that they collect large amounts of versatile individual-level data on students that can be used for learning analytics and optimization of learning processes (e.g. Romero et al., 2008). Since large sets of rich learning data exists today, researchers have been able to examine a variety of different analyses on learning behavior, such as course outcome prediction models.

It is common knowledge, that understanding data and preprocessing it comprise the largest portion of time in the data mining and analysis process. There are oftentimes no exceptions when it comes to learning analytics – LMS data is not exactly designed for data mining and analysis. The main focus of LMS's is to support teaching and learning, which oftentimes means that exporting, cleaning and handling data hasn't been a focus of the developers and thus hasn't been made easy for users. One of the reasons for the difficultness of manual LMS-data handling is the vast quantities of data that these systems generate on a daily basis – personal information, academic results and all logs data related to courses (Romero et al., 2008), tracked with timestamps throughout the course. Moreover, Dyckhoff et al. (2012) noted, that LMS data is oftentimes "incomprehensible, poorly organized, and difficult to follow, because of its tabular format." As LMS-data preprocessing has become such a complex and time-consuming task, researchers such as Krüger et al. (2010) have responded by providing data models to ease the analysis and data mining of educational data.

Other solutions for decreasing the amount of manual data preprocessing work in learning analytics have also been developed. As learning analytics has become an interest of academic administration, dedicated information systems for learning analytics have been developed, and researchers such as Nguyen et al. (2020) have studied and found important principles regarding the design of those learning analytics information systems (LAIS). One of the main principles introduced by them for the design of LAISs, is the Principle of information interoperability. It means that LAISs should enable integration with all educational data systems such as LMSs, without resulting in any detectable complications. However, as LMS-data is oftentimes incomprehensible and poorly organized, integration challenges may appear.

In Daniel's (2015) research on the opportunities and challenges of big data analytics in higher education, data integration arises as one prominent challenge. The study pointed out that educational data is oftentimes stored in systems which are managed by different departments. A university may simultaneously manage a large number of different data warehouses for storing different data. For instance, student demographic information may be stored in a different institutional system than LMS-data, and a different system than official grade or degree data, or student exchange data. Thereby, whenever these data need to be combined for learning analytics, a significant amount of manual work is required. Moreover, according to the research, data sets from different systems may exist in very different formats – structured and unstructured – and may require a lot of data cleansing before performing data integration.

Another principle suggested by Nguyen et al. (2020) for the design of LAISs is the Principle of information anonymity and protection. When implementing learning analytics using educational data, extracted from educational data systems, this is one of the main concerns for the analysis performer and the educational institute. All data related to students and their behavior stored in educational data systems is personal data, which should be always protected to prevent abuse. Laws and regulations, such as data privacy acts ensure such protection of students' personal data, but under certain circumstances, such as clear research purposes, under the condition that the data will be handled transparently and purposefully, consent to the analysis of student data may be given (Dyckhoff et al., 2012). Thus, performing learning analytics requires legal actions in the initiation stage of the project, and consent to data access and analysis may not always be granted.

Research on the implementation challenges of learning analytics and educational data mining today, exists in quite low numbers. However, the few studies that already touch on the challenges of implementing learning analytics, do so quite limitedly and from very specific viewpoints, not focusing on a comprehensive understanding of all challenges related to a data-mining project in the learning analytics context. Literature today also lacks reflexive case studies, which are designed to identify the real-life obstacles in each step of the implementation project. Thereby, a clear research gap exists, that I intend to fill with this research. As some implementation challenges are already recognized in previous studies, solutions to these challenges have not been proposed. The literature that exists, however, provides a great base for

examining the challenges even further, and suggesting proposals for university faculties to overcome these challenges.

3. Methodology

This research is a reflexive research, which analyses the challenges of implementing a course completion prediction model, through a real-life case executed for Aalto University School of Business's marketing courses. This research not only addresses the challenges of implementing the learning analytics project, but also discusses insights on how to overcome these challenges or even avoid these challenges from occurring. I document the obstacles of the modelling project throughout the project timeline, through each stage, all the way from applying for data access approval to running the actual prediction model. I conducted the reflexive case study by writing down the experienced challenges and realizations throughout the project and by analyzing them with insights from previous literature on learning analytics.

I selected the methodology of a reflexive case study, because it is the most insightful approach considering the type of findings that can be made – this research methodology allows for real, experienced challenges of the modelling process to stand out, and thus can provide helpful insights and guidelines for university managers and learning analysts. As this methodology allows the whole project life cycle to be documented in detail, it can help university officials and educational managers pinpoint the challenges in their learning analytics, bring efficiency in their learning analytics processes and improve the quality of their analyses, which then can lead to improved learning of students. Moreover, there's a clear research gap in studies on the challenges of implementing a higher education course completion prediction model. Learning analytics-related studies that address how to improve the implementation of learning analytics projects do exist (see Daniel, 2015; Romero et al., 2008), but do not focus on identifying the challenges comprehensively throughout the implementation processes, and they are not conducted in the context of course completion prediction modelling, nor conducted as reflexive case studies. This reflexive case study on course completion prediction modelling however, covers all of those requirements, and thereby fills a contextual research gap.

4. Findings

To create accurate course completion prediction models, which can predict dropouts across multiple courses, extensive amounts of data from those courses, including individual-level student characteristic and behavior data is needed for variable construction, as well as course-level information in some cases. To examine the challenges of implementing a course completion prediction model, I conduct a case study, in which such a model is created for Aalto University's marketing courses. In section 4.1., I will shortly introduce the case background, and in 4.2. and 4.3. I will thoroughly review the process of implementing the project from beginning to end, simultaneously analyzing the challenges that occur in each stage. Moreover in 4.3., I present the model results.

4.1. Case background

The aim of the case study was to create a prediction model that included data of 14 optional, blended bachelor's level marketing courses, that ran during the academic years 2018-2020 in Aalto University, to predict course dropouts of students in high accuracy across those courses. Those 14 courses included multiple occurrences of six marketing courses, some of which were organized only once, some twice and some three times during the chosen time. The project was conducted in collaboration with Aalto University – the university officials took care of the legality of the project, hand-picking data and integrating data, whereas the rest of the project, including data pre-processing and running the model, was executed by me.

Due to large difficulties in data pre-processing, after all, the model was created using data of only one marketing course targeted for bachelor's students, "Marketing Analytics" from the spring 2019. There were 60 students in the course and the dropout rate of the course was 20% (12 students). I created the model with the binary logistic regression technique in SPSS. The aim was to conduct a model which included three types of data of all 14 courses; LMS-based student behavior data such as all log-ins to course sites, student demographic information such as the students' level of studies, and course-level information such as the difficulty level of the

course. However, due to data integration difficulties, all demographics data had to be eliminated. Moreover, since the final model was ran only for one course, no course-specific information was needed as variables. Thereby, the LMS-data was the only data used in the model. The LMS-data was extracted from MyCourses, which is a Moodle-based learning environment for Aalto University's courses.

4.2. Collecting and pre-processing student data

The project began with seeking approval from the university's learning services. First of all, since the university officials were going to put efforts and resources in the project too, the approval for conducting this type of research was needed, and the approval for accessing loads of student data was needed. The section 4.2.1. will review the beginning of the project and acquiring access to student data. After the approvals were given, the needed data was handpicked by university officials from multiple data systems and integrated together. The section 4.2.2. will dive deeper into the data extracting and integrating stage. The final stage before running the model is data pre-processing, which includes familiarizing with the data, cleaning the data and creating variables. The data pre-processing stage is reviewed in section 4.2.3.

4.2.1. Getting approval to access student data

The project began by contacting university faculty of Aalto University and beginning discussions about conducting the learning analytics project from the university's student data. As the university's strategy aims at developing data-driven activities and using analytics, the idea of developing a course completion prediction model received a very warm welcome and its benefits to students' learning were well understood. The starting point was therefore great for this research, as the university was fully onboard with the project. The prominent issue in getting the approval to data access, however, was data privacy, especially as this type of project which required tons of students' personal data to be handed over to researchers had not been

conducted before in Aalto University – and the same goes to many other universities today. As data privacy is one of the top priorities for educational institutions when it comes to handling data, this created large quantities of work for the university's data privacy lawyers, as they were faced with finding a way to enable giving data access to the research, in a lawful and ethical manner.

Luckily, the university lawyers were able to execute the complicated task and provide a plan for how data privacy could be achieved. The university officials agreed to anonymize the data as well as possible, and I agreed to treat it transparently and purposefully, and minimize the possibility of data misconduct. To ensure the ethicality of the project, necessary legal forms were filled where the research process and purposes of the research were explained, and one was sent over to the board of research integrity, which is the university's organ for evaluating the ethicalness of research projects. After a month was spent on this preparatory work, and almost two months more – partially due to the board of research integrity's summer vacations – were spent on waiting for a response, the verdict came. The research had gotten the green light, and the consent to data access was given.

In this project, the university lawyers were faced with a complicated task of finding out a lawful and ethical way to give the research an access to large quantities of student data. When a research plan requires students' personal data to be handed over from the university to the researcher, multiple workdays are invested in research, evaluating the lawfulness and ethicalness of the research, and paperwork. Moreover, as this learning analytics project was one of the first ones of its kind in the university, thorough planning was needed to complete the tasks. At this stage of development in learning analytics, the specific issues of research needs require careful weighing on a case-by-case basis. For these reasons, the initial schedule expectation I had of this project stage – only a couple of weeks – turned out to be too optimistic, and the project schedule had to be altered.

4.2.2. Extracting and integrating student data

As permission for utilizing student data for research purposes was given, the university's learning services extracted, anonymized, integrated and handed over the data according to the research's needs. This stage of the project also imposed a rather heavy workload on university officials, considering the vast quantity of data, extracting it from two different data systems, and the university staff having to perform data integration for all 14 different courses. The student demographics data picked by one university staff member, had to be integrated with the LMS-data, which was picked by another one. Both of these data were initially located in different data systems. Also course-level information was added to the data, by manually collecting it from information on the course websites, to embody the course-level differences such as course difficulty and course average grade.

The original data thus contained three types of information; two types of individual-level information – student demographics and student behavior, and course-level information:

1. **Student demographics** were extracted from a separate institutional data system and included general student information: the start day of studies, the student's locality status (exchange/local student), student's level of studies (master's/bachelor's) and the major the student is studying.
2. **Student behavior during the course** was extracted as LMS activity data, which included all student activities inside the LMS such as all incidents and timestamps of user logins, opening lecture slides or assignment instructions, opening/writing to discussion forums, and submitting assignments.
3. **Course-level information** on the other hand contained course-specific information about each course in the data set. The information were the following: difficulty level of the course, course average grade, number of course pre-requisites, and language of the course. The course-level data was gathered by hand, from public information on each course's websites.

Even though the data integration was executed successfully and with great precision by the university's learning services, the quality of the data was incomplete. A large portion of the student demographics data that was extracted, was protected due to data privacy issues and thus only error signs showed in most columns which should've contained information. This data loss problem was a large obstacle from the modelling's perspective. Even though this problem

concerned only the students which had not passed the course at hand, whereas all needed information existed at least for all passed students, a course completion prediction model couldn't be built with information that existed for only the students in the data that had not dropped out of the course. Thereby, all demographics-related information had to be eliminated and not used in the model. Unfortunately, this meant a potentially less accurate and less useful model.

4.2.3. Data pre-processing

Data pre-processing included getting to know the 14 datasets, cleaning the data, constructing the same variables for each data set and finally bringing the data sets together as one. Getting to know the data was quite time-consuming, as each course looked very different data-wise. The LMS course-sites of each course had been constructed in very different ways, and consequently, the information on student behaviors were found in different places in each data set. Cleaning the data on the other hand was faster. It included deleting all instances of students from the data that had never attended the course, also deleting all course faculty instances, and all columns in the data, which are known to include partially missing data, like all demographics data, as previously explained. Also, I had to eliminate all group assignment-related data, since typically only one person in each group submits the assignment, which data-wise, falsely looks like the others didn't complete the assignment at all. Moreover, no data existed on which student was in which group. Constructing variables from the raw data on the other hand, was the more challenging and time-consuming task. This stage involved overpowering challenges which lead to having to eliminate all courses from the model, except for one. Thereby, the 14 data sets were never brought together as one.

The main obstacle in data pre-processing was creating similar, comparable variables from data sets of courses, which were so significantly different from each other. First of all, the courses differed substantially in how the course-sites were built in the LMS, which made building variables a very manual work. In the university's LMS, just as in most LMS's, the faculty of each course gets to construct their course pages as they wish and place each item however suits their courses the best. For instance, the assignment instructions could be found in numerous

different places in the LMS course sites – in the front page of the course-sites, in the assignments-item, in the course slides, or in the assignment submission boxes. It appears there were no common instructions for course faculties to building course-sites, or at least they were not followed by course instructors, and thereby the course sites are structurally so diverse.

From data mining perspective, this is a troublesome starting point, as for each course and for each different variable, data miners have to do research on where each item locates in the course sites, before it's possible to begin variable construction. This on the other hand, makes it impossible to automate variable construction – it has to be done manually for each separate course, which is very slow and inefficient. If instead the current approach to building course-sites, all course-sites were built in a similar, standardized way, every item could be found in the same place in the data, and thereby automated variable construction could be possible across all courses.

In addition to the course-site structures, the courses differed from each other in many other ways too, which made the efforts of building comparable variables, almost impossible. They differed in the kind of tasks that were required from students, how the course workload was divided between different tasks, and how on some courses assignments were optional but some mandatory. For instance, one course might require an assignment to be submitted each week, as another course requires only one during the whole course – or one course may require a number of mandatory readings, while others don't at all. The large differences between the tasks required from students, can cause the most predictive variables to differ significantly between courses, and thus make it harder to build one joint model. For these reasons, training a model based on data from 14 very diverse courses, will result in much worse prediction accuracy, than training a model just for one course at a time. Thereby, I made a decision that the model will be conducted for one course only, "Marketing Analytics". As the model was to be conducted for one course only, the course-level data were not needed anymore, and thus I eliminated it at this point.

Another reason for narrowing down the course list was the class imbalance problem: data on some courses simply didn't include enough or at all students who had dropped out. To implement a good course completion prediction model, which is ultimately based on the differences between the dropped and the passed students, it's important that the data includes a

fare share of students who have dropped out, so the model gets to learn their behavior accurately. The class imbalance problem is especially prominent in course completion prediction modelling of higher education, as not very many students commonly drop out of courses in Finland, due to the government's motivating study allowance limitations.

As in "Marketing Analytics" the completion rate was 80%, meaning that the share of dropouts was 20%, the class imbalance problem was existent, but not eminent. However, classification techniques typically give most accurate results when the class division is approximately 50%/50% (Verbeke et al., 2012), so an oversampling method could've been tested for evening out the class distribution – a method in which the instances of dropouts would've been multiplied in the data until the share of dropouts and completers is 50%/50%. However, I didn't find this necessary, since as is reported in part 4.3., the model received great prediction results with the original class distribution.

Even though due to data inconsistency problems I had to narrow down the data to cover just one course, more data challenges emerged concerning the final data set – the data set of "Marketing Analytics" was missing some LMS data. As pre-processing of this course data began, I realized that no discussion forum data existed in the data. The reason was, that the group discussions had been organized in a separate platform Slack, not in the university's MyCourses LMS. Integrating Slack data to the MyCourses data at this stage would've been too troublesome, as the MyCourses data had already been anonymized, which made integrating impossible. Thereby I made the decision to leave out all discussion forum behavior from the model. As the social behaviors of students which the discussion forum data captures, had been found predictive by many previous course outcome prediction studies (e.g. Macfadyen & Dawson, 2010; Taylor et al., 2014; Yang et al., 2013), not being able to use this data unfortunately lowered the possibility of high prediction accuracy.

The missing data problem that occurred, stems from an institution-wide practice; course faculties can organize their courses quite freely, without boundaries which would prohibit organizing discussions or any other tasks on third party applications. If all courses were organized in a way which enables the data collection of each and every course-related student behaviors, prediction models could perform much more accurately. Even though excluding the discussion forum data from the model of the "Marketing Analytics" course was potentially

harmful to the model, an accurate model could still be achieved. If the model had included all 14 courses however, which all – in worst case scenario – had involved a missing data problem, it could've been very detrimental to model accuracy. If course faculties aren't forced to organize the course in a way which collects all possible data within the LMS, the missing data problem will continue to occur, and creating highly accurate prediction models for courses will remain difficult.

4.3. Running the model

I created the prediction model with a binary logistic regression technique in IBM's SPSS. Logistic regression is a well-known classification technique, which is very much used for instance in marketing, as marketers try to predict customer churn. In course completion prediction, the technique aims to classify if a student will or will not complete the university course. In logistic regression, the probability of success, or in this case probability of student retention, p, is constrained to lie between 0 and 1. As I ran the prediction model for student retention, based on the value of p, each student was categorized by the model in to either the group of churned or retained. If the estimated probability of retention (p) is greater than 0.5, the predicted value of Y is set to 1, meaning that the student is likely to be retained. If the estimated value of p is less than 0.5, the predicted value of Y is set to 0, meaning that the student is likely to churn (Malhotra et al, 2012). After I had constructed the final data set, modelling with SPSS was very simple and no significant challenges occurred in the actual modelling stage.

I set the binary variable "Course completed" as the dependent variable of the model. The independent variables on the other hand, were chosen based on an iterative modelling process, where the model was ran over and over with different combinations of variables. First, I chose the combinations of variables by just randomly picking them, and then holding on to the variables which stick out as the most predictive ones. Moreover, the significance levels of the variables were considered in this iterative modelling process. The modelling process was rather fast, when compared to previous stages of the project. As SPSS returns the model immediately after running it, the process was fast, and the final model was reached within just one day of iterative modelling.

All 31 variables that I tested in this modelling process, were largely chosen based on previous studies, which had found certain variables to be predictive in course outcome prediction modelling. These include activity metrics such as course page and lecture slide opens (see Cohen, 2017; Gitinabard et al., 2019) and assignments-related variables (see Halawa et al., 2014; Macfadyen & Dawson, 2010; Taylor et al., 2014). As many first week's activities were found clearly predictive in a student performance prediction study by Jiang et al. (2014), variables embodying the first week's behavior were added. Moreover, I added variables describing the maximum number of a certain behavior in a day or a week, and the weekly drop of a certain behavior, since I hypothesized those variables to have predictive power. I also added grade user report and course syllabus related variables, and tested them during the modelling process, I estimated that they may have high predictive power, too. The list of variables constructed from the data is presented in Table 2.

LMS item	Variable
Course page	Amount_of_course_page_opens_before_course_beginning
	Amount_of_course_page_opens_during_first_week
	Max_amount_of_course_page_opens_per_day
	Amount_of_course_page_opens_per_active_week
	Max_weekly_drop_of_page_opens
	Avg_weekly_drop_of_page_opens
	Week1-2_drop_of_course_page_opens
	Week2-3_drop_of_course_page_opens
	Week3-4_drop_of_course_page_opens
	Week4-5_drop_of_course_page_opens
	Week5-6_drop_of_course_page_opens
Assignment instructions	Amount_of_assignment_instructions_watched_during_first_week
	Max_amount_of_assignment_instructions_watched_per_day
	Amount_of_assignment_instructions_watched_per_active_week
	Max_weekly_drop_of_assignment_intructions_watched
	Avg_weekly_drop_of_assignment_intructions_watched
Lecture slides	Amount_of_lecture_slides_checked_during_first_week
	Max_number_of_lecture_slides_checked_per_day
	Amount_of_lecture_slides_checked_per_active_week
	Max_weekly_drop_of_slides_opened
	Avg_weekly_drop_of_slides_opened
Grade user report	Max_number_of_grade_user_report_checked_per_week
	Max_weekly_drop_of_grade_user_report_checked

	Avg_weekly_drop_of_grade_user_report_checked
Assignments	Amount_of_assignments_submitted Amount_of_assignments_submitted_not_on_time_or_at_all
First assignment	First_assignment_submitted_on_time First_assignment_submitted_late_or_not_at_all
Syllabus	Amount_of_syllabus_watched_before_course_beginning Amount_of_syllabus_watched_during_first_week
Announcements	Amount_of_announcements_checked_during_first_week

Table 2. List of independent variables constructed from raw data

The model with the combination of variables that produced the best prediction accuracy and best statistical significance was found through an iterative testing process, during which the model was ran countless times over and over with different combinations of independent variables. After tens of rounds of modelling, I had found the best model. The model's total prediction accuracy was 93,3%. Out of all 12 students who dropped out during the course, the model was able to correctly predict 9 as true negatives. The remaining three the model could not trace based on the given variables, so the model was able to predict 75% of dropouts correctly. Out of all 48 students who completed the course on the other hand, the model was able to predict 47 as true positives, and predicted only one falsely, resulting in to 97,9% prediction accuracy amongst course completers. Overall, the model was able to classify 56 students out of 60 correctly. Consequently, the overall prediction accuracy is very high, especially when considered how many possibly important variables had to be eliminated due to problems with data quality. The prediction accuracy results are presented in Table 3 and the final predictors, their coefficients, standard errors, odd ratios and statistical significances in Table 4.

Observed		Predicted Course_completed		Percentage Correct
		No	Yes	
Course_completed	No	9	3	75,0 %
	Yes	1	47	97,9 %
Overall Percentage				93,3 %

Table 3. The course completion prediction accuracy of the model

Independent variable	B	S.E.	Sig.	Exp(B)
Amount_of_assignments_submitted	2,98	1,53	0,05	19,75
Amount_of_syllabus_watched_during_first_week	-2,25	1,61	0,16	0,11
Max_number_of_lecture_slides_checked_per_day	2,99	1,48	0,04	19,83
Max_amount_of_assignment_instructions_watched_per_day	-1,23	0,60	0,04	0,29
Avg_weekly_drop_of_course_page_opens	-0,63	0,41	0,12	0,53
Avg_weekly_drop_of_assignment_intructions_watched	2,51	1,93	0,19	12,33
Constant	-11,60	7,33	0,11	0,00

Table 4. Summary of logistic regression analysis for variables predicting course completion

The preferable significance level for this model was 0,05, meaning that each variable with significance $\alpha>0,05$ should've been discarded from further analysis as they have too high probability of rejecting the null hypothesis (Malhotra et al. 2012) which is, that the variable has predictive power in course completion. As Table 4 shows, the significance levels of some variables in this model however, exceed the wished limit, up to being significant only at $\alpha>0,2$ level. As explained previously, due to large problems in data quality, I had to discard a notable amount of data, and thus more variables could not be constructed and tested. Thereby, this model doesn't include all features that might be important predictors of course dropouts, which leads to weaker significance levels. Another factor causing slightly weak significance levels could be the low sample size (60 students) and low portion of course dropouts (12 students). However, after a long iterative modelling process, this model has the best combination of high prediction accuracy and low significance levels, that the current data set was able to produce.

Odd ratios are used to describe how much the probability of retention is increased or decreased by one unit increase in the variable, or in other words, how successful the variable is in predicting churn (e.g. Hawkins et al., 2013; Y. Lee, 2018; Wladis et al., 2014). As Table 4 shows, the final model included six independent variables – three of which notably increased the probability of retention; amount of assignments submitted, maximum number of lectures slides checked per day, and average weekly drop of assignment instructions watched. For one unit increase in a student's maximum amount of lecture slides watched per day, the probability of retention increased by 19,8 times. For the three other variables on the other hand – amount

of syllabus watched during first week, maximum number of assignment instructions watched per day, and average weekly drop of course page opens – a unit increase in the variable decreases the probability of retention. The log odds of each variable are interpreted and analyzed in Table 5.

Independent variable	Log odds	Effect on completion probability	Interpretation	Analysis
Amount_of_assignments_submitted	19,75	Increasing	The more assignments the student has submitted, the more the student is committed to completing the course.	The bias of sunk costs exists. As students have spent a lot of valuable time in completing assignments, they are less likely to waste all those hours and drop out.
Amount_of_syllabus_watched_during_first_week	0,11	Decreasing	The more the student has opened the course syllabus during first week, the less likely the student is to complete the course.	As a student reads the syllabus multiple times, which includes all course requirements, they easily realize that they truly don't have the resources to complete the course.
Max_number_of_lecture_slides_checked_per_day	19,83	Increasing	The more a student opens the lecture slides per day, the more likely they are to complete the course.	If a student opens lecture slides multiple times a day, it shows their committment to the course.
Max_amount_of_assignment_instructions_watched_per_day	0,29	Decreasing	The more a student opens assignment instructions per day, the less likely they are to complete the course.	If a student has to watch the assignment instructions over and over again, multiple times a day, it shows a sign of incapability to complete the assignment, which can lead to course dropouts.
Avg_weekly_drop_of_course_page_opens	0,53	Decreasing	The more the amount of course page opens drops each week, the less likely the student is to complete the course.	If a student opens the course page less and less each week, they are showing less committment to the course, and thus may drop out.
Avg_weekly_drop_of_assignment_intructions_watched	12,33	Increasing	The more the amount of assignment instructions opened drops each week, the more likely the student is to complete the course.	If a student opens the assignment instructions less and less each week, it means they have opened assignment instructions multiple times in the beginning of the course, meaning that they are aware of the future workload, prepared for it, and thus more likely to complete the course.

Table 5. Summary of the interpretation and analysis of each independent variable's log odds

The reason behind conducting course completion prediction models usually is that the students at-risk of dropping out of the course could be discovered early on, so they could be targeted with retention interventions such as tutoring and thus prevented from dropping out. As the prediction model in this study was able to predict only 9 of the actual 12 churners correctly, targeting only the top-12 most likely churners with retention interventions, would not have

sufficed to capture all of the actual churners. To examine the accuracy of the model further, I calculated the lift of the model. Lift is probably the most used way to measure the accuracy of course completion prediction models (see Burez & Van den Poel, 2007; Neslin et al., 2006). Thus, Chart 1 (Cumulative Gains Chart) and Chart 2 (Lift Chart) were created. In the context of course completion prediction, lift helps to understand the effectiveness of targeting the students predicted as most likely to not complete the course.

After I sorted the students in ascending order by their high dropout figure, or low course completion figure, which is calculated with the coefficients of the model's predictive variables, I could create Chart 1 (Cumulative gains chart). In this chart, the lift curve shows what cumulative percentage of actual dropouts could be reached by contacting x% of the predicted dropouts in their sorted order. The baseline on the other hand shows the cumulative percent of actual churners reached if no prediction model is used. The chart reveals that to reach for instance 92% of all actual churners, only 30% of the students with the lowest predicted course completion figures need to be contacted with retention interventions, whereas by using no model at all, and thus contacting a random 30%, only 30% of the actual churners could be reached. This implies that the prediction model offers great efficiency for targeting students based on the model. Course faculties could offer personalized tutoring or other retention interventions for the top 30% with lowest predicted course completion figures in the course (60*30%=18 students), and thus possibly prevent the dropouts of 11 students (12*92%=11). Thereby, all but one actual churner could be reached and possibly retained, by contacting only 30% of the class.

Chart 2 (Lift Chart) on the other hand shows us how much more likely it is to contact actual churners using the prediction model's result, than if we contact a random sample of customers. For instance, by contacting 30% of the students predicted as most likely to dropout, 3,07 times more actual churners are reached than if a random 30% sample of the students is contacted. The cumulative lift curve thus shows the difference in effect between using a model and not using any model (e.g. Burez & Van den Poel, 2007; Hwang et al., 2004; Neslin et al., 2006). Clearly, while contacting any percentage of students with retention interventions, the model results are able to outperform using a random sample. In the customer churn prediction context of businesses, which may include some very costly targeted retention campaigns with discounts

and offers, the higher the lift of the customer churn prediction model, the more profitable the company's churn interventions will be (Burez & Van den Poel, 2007).

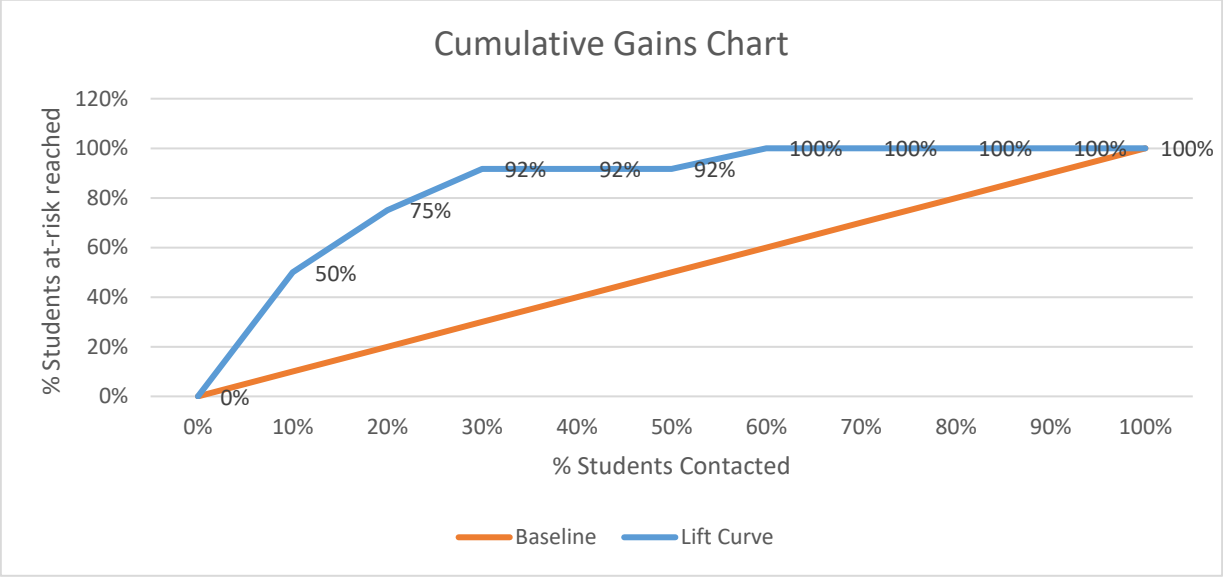

Chart 1. Cumulative Gains Chart

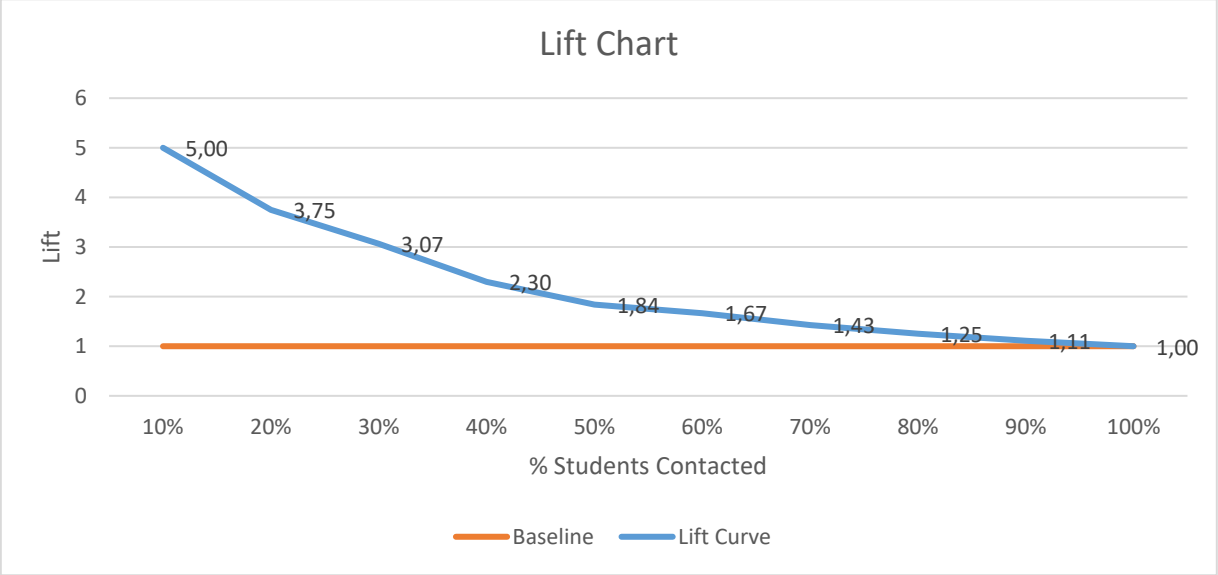

Chart 2. Lift Chart

5. Discussion

The motivation for conducting this study was that the clear research gap in the challenges of implementing learning analytics in institutions of higher education. More specifically, the challenges of such implementation had not been researched through a reflexive case study which in fact analyses and reflects on the challenges that occur in a real-life learning analytics project. This study successfully fills that research gap by introducing a course completion prediction model project, which is conducted by constantly writing down the challenges that occur during the process, analyzing them and developing solutions to overcome them. The study does not just focus on solving a specific challenge with the implementation process but comprehensively examines all of the obstacles this experienced project came across, considering challenges with the university's stakeholders and processes, the course structures and course-sites, the data and the modelling technology. The most prominent challenges which occurred were related to data quality, since a lot of data had to be eliminated due to missing data problems, and the modelling data had to be reduced in many ways, resulting into a small sample of only 60 students.

Another research gap filled by this research is related to the challenges of the data privacy and data access process when implementing learning analytics in a higher education institution. Learning analytics literature today does not yet identify the challenges in the process stage. This research however, provides detailed information concerning the experienced data privacy process, highlighting the heavy workload of lawyers and data privacy officials, and the additional efforts that go in to research and preparatory work if the university is dealing with a completely new type of data privacy case. This paper also recognizes, that if a research requires vast amounts of student data to be handed over from a university, the researcher should familiarize with the institutions' policies and inner processes beforehand, in order to make an executable research plan and a realistic project schedule.

Learning analytics researchers today document the methodology of the data collection process very thoroughly, usually describing how the data was extracted from data systems and integrated with other data. However, to keep research papers compact, small challenges which occur in the data extracting and integration stage are oftentimes not discussed in the papers, if the focus of the paper is not the challenges themselves. There are however, some previous studies which do

observe these obstacles (see Daniel, 2015; Nguyen et al., 2020), especially from the viewpoint of data interoperability and consequent loss of data, but their findings were not the result of a reflexive case study. This research paper on the other hand supplements the findings of those papers, revealing the real-life consequences of loss of data, based on an experienced learning analytics project, conducted by reflexive case study methods. Due to loss of data, all of the student demographics data had to be fully discarded in this project, which potentially an impact on the model results. Thus, this research was able to produce insightful information on the problems of data extracting and integration in learning analytics by filling the research gap.

This study concentrated on the problem of low data quality in the implementation of learning analytics, to better understand the obstacle and to enable overcoming it in the future. Daniel's (2015) argument about low quality data and how the lack of standardized measures and indicators make comparison difficult, turned out to be one of the main challenges in this reflexive case study. This study, however, was able to identify a source of that low data quality problem, which previous literature had not introduced – the severe diversity of courses and course-site structures. The heterogeneity of course-site structures displayed as low-quality data and made creating variables very manual work, precluding the automation of analytics and effortless comparisons between courses. Also, I found another challenge in data pre-processing of this study, which has not been introduced as an obstacle in learning analytics literature yet – a missing data problem caused by the decision of course faculty to not arrange group discussions within the LMS platform. Overall, the research gap of the challenges of data pre-processing in learning analytics implementation, were filled in this research.

As literature today already covers course outcome prediction studies in the MOOC context to a large extent, the situation is quite different in the blended learning context. In this study, although from a reflexive case study perspective, I conducted a course outcome prediction model in the blended learning context of higher education and pinpointed the project's challenges in doing so. The largest challenge I discovered concerning the context of blended learning – also mentioned by previous researchers (see Gitinabard et al., 2019), was the inability to collect data from students' learning in face-to-face classes and other offline learning activities. Luckily, the one course chosen for the modelling had required a lot of online activities from students, which were captured by the LMS system and thus enough informational data was

collected throughout the course timeline. Therefore, it was possible to produce a model with great prediction accuracy. Had the course required a low number of measurable online activities and a high number of unmeasurable offline activities, the prediction accuracy results could have looked much different.

5.1. Implications

During the implementation of the modelling project, different kinds of challenges and obstacles occurred in varying stages of the project. More specifically, the challenges occurred during the following work tasks; 1. Getting approval to access data, 2. Extracting and integrating data and 3. Data pre-processing, which consisted of getting to know the data, cleaning the data and constructing variables. During the actual modelling stage however, no obstacles occurred, as I had already pre-processed the data to its final form and the modelling technology worked as expected. To implement course completion prediction modelling projects, and other learning analytics projects in a higher education setting successfully and without large complications, the discovered obstacles need to be overcome within higher education institutions by developing their internal processes, course-site structures data systems. These challenges are summarized and analyzed in the next paragraphs, and possible solutions to university faculties are suggested for overcoming these challenges.

Proposition 1. Create a standardized process for applying for data access

The projects first stage; data privacy and getting approval to access data, included multiple meetings with the university's learning services officials and lawyers, and loads of paperwork concerning data privacy. As data privacy is a top concern for universities in this type of projects and extreme carefulness is required of legal preparatory work, heavy workloads and time consumption in the data access stage are inevitable. Currently, each separate student data access request from researchers requires precise case-by-case processing in Aalto University, partially due to the unique features of each request. However, if learning analytics projects are to be conducted in higher frequency in the future, a new internal process could be created and

developed to support the data privacy-related work and to lighten the workload of university officials at least to some extent.

Thereby, I propose that university managers develop a standardized process for applying data access for learning analytics projects. For instance, clear criteria for the project's approval approved by the university lawyers should be set in advance, and certain officials should be named as responsible of reviewing the project. If routine-like tasks in the process can be identified, university officials can look further into automatizing some of those tasks and utilizing scalability. Thereby, both the workload and the life-cycle of the process could be improved. Aalto University has already recognized the need for such defined processes and policies regarding learning analytics and personal data. The university has already taken a leap forwards by developing a learning analytics policy, which defines the roles, responsibilities and access to learning and teaching data in order to support learning, teaching and university decision making by the means of learning analytics.

Proposition 2. Adopt a learning analytics information system to ease data integration

The obstacles which occurred during the stages of extracting and integrating data, were the result of data quality challenges. As universities oftentimes store data in multiple different data systems managed by different departments, and as LMS data oftentimes have poor interoperability with other data (Daniel, 2015; Nguyen et al., 2020), problems with data integration are likely to occur. Data integration in the prediction modelling case study, was implemented manually by the university's learning services. Data interoperability resulted into a significant quantity of missing data, which from the model's viewpoint, was harmful. To result into high data quality, data needs to be integrated without resulting in to missing data, which could be achieved by adopting a learning analytics information system (LAIS) or a similar system in the future.

Many researchers describe LAISs as distinct, standalone systems which are designed to integrate and interpret data from multiple educational data systems, and thus help analyze it (see Dyckhoff et al., 2012; Nguyen et al., 2020). A LAIS should integrate all relevant data from different data systems successfully, and not result into missing data. Moreover, it could significantly

accelerate data pre-processing times, as it is able to combine and display actionable information of students, which helps for instance with constructing variables. However, finding and adopting a standalone LAIS service could be an impossible task for now, since currently there are no such solutions on the markets that fill all of the mentioned requirements.

Proposition 3. Harmonize course and course-site structures to enable automated variable construction

As the courses differed from each other by a variety of different aspects, and as the course sites differed from each other significantly, getting to know the data, cleaning the data and constructing variables had to be implemented manually for each separate course. This was very time-consuming, especially as every course-site was structurally unique and thereby every data set contained items locating in different columns and by different names. To avoid this problem in the future, from course completion prediction modelling's perspective, I propose that all courses would be structured at least slightly more similarily. If the courses used in learning analytics resembled each other to a high degree and they would be comparable, it would be more likely that learning analytics could be implemented in an automated way, across all courses. Moreover, the course-site structure in LMS's should be standardized, so that all items would be found in the same order in all course pages. Consequently, if in addition to course structures, also course-sites were harmonized, it would be possible to completely automate the data preprocessing stage of learning analytics across courses. These changes to courses and course-sites however, would require an institution-wide change and a project team to implement the change.

Proposition 4. Organize courses in a way which supports data generation

One challenge that occurred while familiarizing with the data, was that some data was not collected at all in the LMS sites of some courses. This was the causation of course faculty using other platforms than the course LMS to facilitate group discussions. Another problem related to uncollected data was already acknowledged in the beginning of the project – as LMS's only collect data from online behavior, offline activities were not captured as data. To overcome

these challenges, universities could organize courses and LMS usage in a way which generates more useful and accessible data. If the course execution allows for it, the course's group discussions could be organized inside the standard LMS, and the usage of outside services could be minimized. Moreover, if course faculties of blended learning courses wish to utilize learning analytics to support their students' learning, standardized ways of collecting students' interaction data during offline activities could be adopted, and thus great value could be added to the learning analyses. For instance, collecting data on students' attitudes and interest levels during lectures could be executed by adopting a compulsory lecture survey, which would be filled at the end of each lecture by each student. If more behavioral factors of students could be collected and analyzed, more accurate learning analytics could be conducted, and the university could support their students' learning to even a higher degree based on analytics.

5.2. Limitations

Since the reflexive research bases on experiences with only one higher education institution, making the sample size quite small, the findings may not include all possible challenges that learning analysts may face when dealing with other higher education institutions. However, the findings of challenges that were identified, are very generalizable, especially in the western world, since the chosen university is a typical western university with high data privacy standards, and data collection and data storage processes typical to higher education institutions of today. Universities and colleges which locate in a country with differing requirements in data privacy legislation, however, may experience different challenges in the data privacy stage when implementing a learning analytics project.

It's important to note that I conducted the model only using data of one specific higher education marketing course in the blended learning context, so the actual modelling results cannot be applied on other types of courses – not even other similar marketing courses, before more research is conducted. The reason for this is the large differences of courses and the consequent differences between student behavior on those courses, which the model is trained upon. The model I created in this study thereby only provides insights of the one specific course, Marketing Analytics in Aalto University. To discover the features of students which best predict student

churn concurrently in all higher education marketing courses, a new modelling project needs to be conducted which includes data of all different types of marketing courses. Moreover, more advanced modelling methods should be explored, to reach the best possible results.

Moreover, the findings concerning the modelling and its predictors shouldn't be applied on other similar marketing analytics courses either, because of the variables in the model were not statistically significant at the $p<0,05$ level. Nevertheless, considering the predictive power and statistical significance, the model was the best one that could be produced with the 31 constructed variables. Since I had to eliminate all demographics data due to missing data problems, the demographic variables that previous literature had discovered to have high predictive power in higher education, could not be tested during the modelling process. Moreover, I excluded all discussion forum data and group assignment data from the model. Therefore, whether a model with higher predictive power and statistical significance could've been reached with those variables, remains unknown.

6. Conclusions

The aim of this study is to identify the real-life challenges of implementing a learning analytics project for a higher education institution in the blended learning context, and to suggest improvement propositions for institutions for overcoming the challenges in their processes of researching and developing learning analytics. Identifying these challenges enables policymakers and learning analysts to make informed decisions when considering implementing learning analytics research and development projects, and especially course completion prediction in their institutions.

First of all, the reflexive case study showed that to implement learning analytics more effortlessly in the future, especially if implemented in high frequency, higher education institutions should create a standardized process for dealing with data privacy and data access. Moreover, it revealed that the process of extracting and integrating student data from different data systems may lead to a loss of data due to data privacy protection or data quality, and thereby could be improved for instance by adopting a learning analytics information system. Thirdly, an

institution-wide harmonization of course and course-site structures needs to be conducted to enable automated variable construction, measurability, and data comparability. The final main finding was, that course faculties may cause loss of data themselves by their decisions regarding how the course is arranged, and therefore, clear organization-wide guidelines should be set, to ensure the collection of all necessary learning data.

One interesting area of further research is the effect of the share of offline activities in a blended learning course to the prediction accuracy of course outcome prediction. Assumably, the prediction accuracy of the model in this study could've improved even further if informational data from the offline activities would've been collected and included in the model as well. Further research on course outcome prediction in blended learning also should look into how well different offline activity variables can predict course outcomes; such as variables describing the attendance on offline lectures, students' level of engagement during offline activities and students' social activities during offline activities.

Before conducting research on those offline variables, however, higher education institutions should look in to to developing and harnessing methods to measure and collect such offline student activity data, since currently such measuring technologies are not extensively in use in higher education, and such data is therefore not widely available. As in this research – due to prominent challenges in data preprocessing – a prediction model was created to apply for one course only, future research directions involve a look into whether by conforming to the improvement proposals of this research, a high accuracy course outcome prediction model can be created which applies on multiple courses in the blended learning context.

7. References

Atchley, W., Wingenbach, G., & Akers, C. (2013). Comparison of Course Completion and Student Performance through Online and Traditional Courses. *International Review of Research in Open and Distributed Learning*, *14*(4), 104–116. https://doi.org/10.19173/irrodl.v14i4.1461

Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to

reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, *32*(2), 277–288. https://doi.org/10.1016/j.eswa.2005.11.037

Burgos, C., Campanario, M. L., Peña, D. de la, Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mijning for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*, *66*, 541–556. https://doi.org/10.1016/j.compeleceng.2017.03.005

Cohen, A. (2017). Analysis of student activity in web-supported courses as a tool for predicting dropout. *Educational Technology Research and Development*, *65*(5), 1285–1304. https://doi.org/10.1007/s11423-017-9524-3

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, *34*(1), 313–327. https://doi.org/10.1016/j.eswa.2006.09.038

Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, *46*(5), 904–920. https://doi.org/10.1111/bjet.12230

Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and implementation of a learning analytics toolkit for teachers. *Educational Technology and Society*, *15*(3), 58–76.

Gitinabard, N., Xu, Y., Heckman, S., Barnes, T., & Lynch, C. F. (2019). How Widely Can Prediction Models Be Generalized? Performance Prediction in Blended Courses. *IEEE Transactions on Learning Technologies*, *12*(2), 184–197. https://doi.org/10.1109/TLT.2019.2911832

Halawa, S., Greene, D., & John, M. (2014). Dropout prediction in MOOCs using learner activity features. In *Proceedings of the European MOOC Stakeholder Summit 2014* (Issue January). http://www.emoocs2014.eu/sites/default/files/Proceedings-Moocs-Summit-2014.pdf

Hawkins, A., Graham, C. R., Sudweeks, R. R., & Barbour, M. K. (2013). Academic

performance, course completion rates, and student perception of the quality and frequency of interaction in a virtual high school. *Distance Education*, *34*(1), 64–83. https://doi.org/10.1080/01587919.2013.770430

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications*, *26*(2), 181–188. https://doi.org/10.1016/S0957-4174(03)00133-7

Ibrahim, Z., & Rusli, D. (2007). Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision tree And Linear Regression. *Proceedings of the 21st Annual SAS Malaysia Forum*, *September*, 1–6. https://www.researchgate.net/profile/Daliela_Rusli/publication/228894873_Predicting_Students'_Academic_Performance_Comparing_Artificial_Neural_Network_Decision_Tree_and_Linear_Regression/links/0deec51bb04e76ed93000000.pdf

Jiang, S., Williams, A. E., Schenke, K., Warschauer, M., & Dowd, D. O. (2014). Predicting MOOC Performance with Week 1 Behavior. *Educational Data Mining 2014,* 273–275.

Joksimovic, S., & Gaševic, D. (2015). Social presence in online discussions as a process predictor of academic performance. *Journal of Computer Assisted Learning*, *31*(6), 638–654. https://doi.org/10.1111/jcal.12107

Jordan, K. (2015). Massive Open Online Course Completion Rates Revisited: Assessment, Length and Attrition. *International Review of Research in Open and Distributed Learning*, *16*(3), 341–358. https://doi.org/10.19173/irrodl.v16i3.2112%0AArticle

Kovanovic, V., Joksimovic, S., Gasevic, D., & Hatala, M. (2014). What is the Source of Social Capital ? The Association Between Social Network Position and Social Presence in Communities of Inquiry. *Proceedings of the Workshops Held at Educational Data Mining 2014, Co-Located with 7th International Conference on Educational Data Mining (EDM 2014)*.

Krüger, A., Merceron, A., & Wolf, B. (2010). A data model to ease analysis and mining of educational data1. *Educational Data Mining 2010 - 3rd International Conference on Educational Data Mining*, *October 2010*, 131–140.

Lee, Y. (2018). Effect of uninterrupted time-on-task on students ' success in Massive Open Online Courses (MOOCs). *Computers in Human Behavior*, *86*, 174–180. https://doi.org/10.1016/j.chb.2018.04.043

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Neslin, S. A., Gupta, S., Kamakura, W., Junxiang, L. U., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Nguyen, A., Tuunanen, T., Gardner, L., & Sheridan, D. (2020). Design principles for learning analytics information systems in higher education. *European Journal of Information Systems*, *00*(00), 1–28. https://doi.org/10.1080/0960085X.2020.1816144

Olivé, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2020). A supervised learning framework : using assessment to identify students at risk of dropping out of a MOOC. *Journal of Computing in Higher Education*, *32*(1), 9–26. https://doi.org/10.1007/s12528-019-09230-1

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers and Education*, *51*(1), 368–384. https://doi.org/10.1016/j.compedu.2007.05.016

Sunar, A. S., White, S., Abdullah, N. A., & Davis, H. C. (2017). How learners ' interactions sustain engagement : a MOOC case study. *IEEE Transactions on Learning Technologies*, *10*(4), 475–487.

Taylor, C., Veeramachaneni, K., & O'Reilly, U.-M. (2014). *Likely to stop? Predicting Stopout in Massive Open Online Courses*. http://arxiv.org/abs/1408.3382

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229.

https://doi.org/10.1016/j.ejor.2011.09.031

Wladis, C., Conway, K., & Hachey, A. C. (2017). Using course-level factors as predictors of online course outcomes: a multi-level analysis at a US urban community college. *Studies in Higher Education*, *42*(1), 184–200. https://doi.org/10.1080/03075079.2015.1045478

Wladis, C., Hachey, A. C., & Conway, K. (2014). An investigation of course-level factors as predictors of online STEM course outcomes. *Computers and Education*, *77*, 145–150. https://doi.org/10.1016/j.compedu.2014.04.015

Yang, D., Sinha, T., & Adamson, D. (2013). " Turn on , Tune in , Drop out ": Anticipating Student Dropouts in Massive Open " Turn on , Tune in , Drop out ": Anticipating student dropouts in Massive Open Online Courses. *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, *11*(December), 14.

Zafra, A., & Ventura, S. (2012). Multi-instance genetic programming for predicting student performance in web based educational environments. *Applied Soft Computing Journal*, *12*(8), 2693–2706. https://doi.org/10.1016/j.asoc.2012.03.054