Decision Support Systems 135 (2020) 113325

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

# Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model

Kristof Coussement[a,b,*], Minh Phan[a,b], Arno De Caigny[a,b], Dries F. Benoit[c], Annelies Raes[d]

[a] IESEG School of Management, 3 Rue de la Digue, F-59000 Lille, France
[b] LEM-CNRS 9221, 3 Rue de la Digue, F-59000, Lille, France
[c] Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, B-9000 Gent, Belgium
[d] ITEC, imec Research Group at KU Leuven, Kapeldreef 75, B-3001 Leuven, Belgium

ARTICLE INFO

*Keywords:*
Learning analytics
Proactive student management
Subscription-based online learning
Student dropout
Logit leaf model
Machine learning

ABSTRACT

Online learning has been adopted rapidly by educational institutions and organizations. Despite its many advantages, including 24/7 access, high flexibility, rich content, and low cost, online learning suffers from high dropout rates that hamper pedagogical and economic goal outcomes. Enhanced student dropout prediction tools would help providers proactively detect students at risk of leaving and identify factors that they might address to help students continue their learning experience. Therefore, this study seeks to improve student dropout predictions, with three main contributions. First, it benchmarks a recently proposed logit leaf model (LLM) algorithm against eight other algorithms, using a real-life data set of 10,554 students of a global subscription-based online learning provider. The LLM outperforms all other methods in finding a balance between predictive performance and comprehensibility. Second, a new multilevel informative visualization of the LLM adds novel benefits, relative to a standard LLM visualization. Third, this research specifies the impacts of student demographics; classroom characteristics; and academic, cognitive, and behavioral engagement variables on student dropout. In reviewing LLM segments, these results show that different insights emerge for various student segments with different learning patterns. This notable result can be used to personalize student retention campaigns.

## 1. Introduction

Online learning refers to educational activities in cyberspace; it has transformed educational markets [1]. Educational content providers, both academic and commercial, use the Internet as a primary content delivery channel [2,3], to leverage its advantages in relation to content diversity, flexibility, scalability, accessibility, and cost effectiveness [4]. Among the various business models, online subscription-based learning platforms (e.g., Datacamp, Coursera, Lynda) are the most popular. They require students to pay a weekly, monthly, or yearly subscription fee to gain access to all course content on the learning platform during that subscription. Thus students can follow their own timeline: begin their learning experience at any point in time, select a personalized set of courses, complete the courses at their own pace, and halt the learning process independently. Yet despite these advantages, online learning suffers high student dropout rates [5,6], which vary between 25% and 90%—significantly higher than the rates for on-campus courses [7–10]. Dropouts are problematic for both students and providers, in that students who drop out suffer diminished confidence and motivation to continue future online learning ventures [11], and online learning providers must worry about course quality and the potential negative effects on their rankings and profits [12,13].

To combat student dropouts, online content providers can turn to predictive analytics, which comprise various methods to learn from historical data and predict a future event [14–17]. In detail, online student dropout predictions constitute binary classification tasks: On the basis of historical student variables, the provider assigns students a binary label that indicates their future learning behavior, namely, drop out or stay [18]. The student variables might include demographics, classroom characteristics, or individual needs, as well as cognitive, academic, and behavioral engagement variables. Tables 1 and 2 list some pertinent research and also reveal some gaps. First, existing research often includes a limited number of predictors that capture only a few determinants of a student's learning behavior, even though online

* Corresponding author.
*E-mail addresses:* k.coussement@ieseg.fr (K. Coussement), m.phan@ieseg.fr (M. Phan), a.de-caigny@ieseg.fr (A. De Caigny), dries.benoit@ugent.be (D.F. Benoit), annelies.raes@kuleuven.be (A. Raes).

https://doi.org/10.1016/j.dss.2020.113325
Received 18 December 2019; Received in revised form 15 May 2020; Accepted 20 May 2020
Available online 26 May 2020

**Table 1**
Literature pertaining to online learning dropout prediction.

| Reference | Year | Online learning context[+] | Data set (number of observations) | Number of variables | Variable Groups | | | | | | | Predictive algorithm[*] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Demographics | Classroom characteristics | Individual needs | Affective engagement | Cognitive engagement | Academic engagement | Behavioral engagement | |
| Kotsiantis, Pierrakeas, & Pintelas [29] | 2003 | SPOC | 1 online course of Hellenic Open University (354 students) | 11 | ✓ | – | – | – | – | ✓ | – | DT, NN, NB, LR, SVM, kNN |
| Lykourentzou, Giannoukos, Nikolopoulos, Mpardis, & Loumos [60] | 2009 | e-learning | 2 online courses, NET and WEB (130 and 63 students) | 9 | ✓ | – | – | – | – | ✓ | – | NN, SVM, PESFAM |
| Balakrishnan & Coetzee [61] | 2013 | MOOCs | 1 online course on edX (29,882 students) | 4 | – | – | – | – | – | ✓ | ✓ | HMM |
| Sharkey & Sanders [62] | 2014 | MOOCs | 1 online course (20,000 students) | 15 | ✓ | ✓ | – | – | – | ✓ | ✓ | RF |
| Amnueypornsakul, Bhat, & Chinprutthiwong [63] | 2014 | MOOCs | 1 online course (36,583 students) | 14 | – | – | – | – | – | ✓ | ✓ | SVM |
| Kloft, Stiehler, Zheng, & Pinkwart [64] | 2014 | MOOCs | 1 online course (20,828 students) | 22 | ✓ | ✓ | – | – | – | ✓ | ✓ | SVM |
| Jiang, Williams, Schenke, Warschauer, & O'dowd [30] | 2014 | MOOCs | 1 online course on Coursera (37,933 students) | 4 | ✓ | ✓ | – | – | – | ✓ | – | LR |
| Whitehill, Williams, Lopez, Coleman, & Reich [65] | 2015 | MOOCs | 10 online courses on edX (245,034 students) | 37 | – | – | – | – | – | ✓ | ✓ | LR |
| Tan & Shao [43] | 2015 | e-learning | e-Learning at Open University of China (62,375 students) | 26 | ✓ | – | – | – | – | ✓ | – | NN, DT, BN |
| He, Bailey, Rubinstein, & Zhang [66] | 2015 | MOOCs | 2 online courses (85,281 students) | 7 | – | – | – | – | ✓ | ✓ | – | LR |
| Koedinger, Kim, Jia, McLaughlin, & Bier [31] | 2015 | MOOCs | 1 online course on Coursera (27,720 students) | 3 | – | – | – | – | – | ✓ | – | LR |
| Chaplot, Rhim, & Kim [67] | 2015 | MOOCs | 1 online course on Coursera | 7 | ✓ | ✓ | – | ✓ | – | ✓ | ✓ | NN, HMM, RF, SVM |
| Boyer & Veeramachaneni [68] | 2015 | MOOCs | 1 online course on edX (235,197 students) | 14 | – | – | – | – | – | ✓ | – | LR |
| Fei & Yeung [69] | 2015 | MOOCs | 2 online courses on Coursera and edX (39,877 and 27,629 students) | 7 (Coursera), 5 (edX) | – | – | – | – | – | ✓ | ✓ | HMM, RNN |
| Xing, Chen, Stein, & Marcinkowski [70] | 2016 | MOOCs | 1 online course on Canvas (3617 students) | 6 | ✓ | – | – | – | – | ✓ | ✓ | BN, DT |
| Qiu et al. [71] | 2016 | MOOCs | 11 online courses on edX (88,112 students) | 34 | ✓ | – | – | – | – | ✓ | ✓ | LR, SVM, FM, LadFG |
| Robinson, Yeomans, Reich, Hulleman, & Gehlbach [72] | 2016 | MOOCs | Online courses by HarvardX (41,946 students) | 38 | ✓ | – | – | – | ✓ | – | – | LR |
| Liang, Li, & Zheng [40] | 2016 | MOOCs | 39 courses on edX (200,904 students) | 34 | ✓ | ✓ | – | – | ✓ | ✓ | – | SVM, LR, RF, BOOST |
| Nagrecha, Dillon, & Chawla [47] | 2017 | MOOCs | 1 course on edX | 14 | – | ✓ | – | – | – | ✓ | – | DT, LR, RF, BOOST_ |

(*continued on next page*)

**Table 1** (*continued*)

| Reference | Year | Online learning context[+] | Data set (number of observations) | Number of variables | Variable Groups | | | | | | | Predictive algorithm[*] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Demographics | Classroom characteristics | Individual needs | Affective engagement | Cognitive engagement | Academic engagement | Behavioral engagement | |
| Al-Shabandar et al. [46] | 2017 | MOOCs | 15 online courses by Harvard and MIT (597,692 students) | 14 | ✓ | ✓ | – | – | – | ✓ | ✓ | DT, RF, SVM, NB, NN, LR, SOM, LDA |
| Wang, Yu, & Miao [39] | 2017 | MOOCs | 39 online courses on XuetangX (120,542 students) | 186 | – | ✓ | – | – | – | ✓ | ✓ | SVM, LR, DT, BOOST, RF, NB, DNN |
| Burgos et al. [73] | 2018 | e-learning / distance learning | Moodle DB (100 students) | 6 | – | – | – | – | – | ✓ | – | LR |
| This study | | Subscription-based online learning | 142 online courses provided on one e-learning platform (10,554 students) | 122 | ✓ | ✓ | – | – | ✓ | ✓ | ✓ | LLM, LR, SVM, NN, DT, RF, BAG, BOOST, LMT |

+ Small Private Online Course (SPOC), Massive Open Online Course (MOOC).

[*] *Predictive algorithm abbreviation:* Bagged Decision Tree (BAG), General Bayesian Network (BN), Boosting Tree (BOOST), Deep Neural Network (DNN), Decision Tree (DT), Factorization Machines (FM), Hidden Markov Models (HMM), K-Nearest Neighbor (kNN), Latent Dynamic Factor Graph Model (LadFG), Linear Discriminant Analysis (LDA), Logit Leaf Model (LLM), Logistic Model Tree (LMT), Logistic Regression (LR), Gaussian Naïve Bayes (NB), Neural Network (NN), Probabilistic ensemble simplified fuzzy ARTMAP (PESFAM), Random Forest (RF), Recurrent Neural Network (RNN), Self-Organized Map (SOM), Support Vector Machines (SVM).

learning platforms involve a vast multitude of behaviors and interactions, resulting in large and rich student data [19]. Second, the five most common online learning dropout predictive models—logistic regression (LR), support vector machines (SVM), neural network (NN), decision tree (DT), and random forests (RF)—all struggle to balance predictive accuracy with comprehensibility, but such features both are critical for targeting students effectively with personalized retention campaigns. For example, though DT assessments can identify different segments in terms of dropout behavior, they typically lack good generalizability. Whereas LR offers a comprehensible predictive model, its performance is not as strong as SVM, NN, or RF. These latter models typically achieve better predictive performance than the DT and LR, but they are nonlinear classifiers and thus cannot explicate heterogeneity among students, so they result in less comprehensible or actionable classification models.

In recognizing these gaps, we pursue three contributions with this study. First, we introduce the logit leaf model (LLM) to an online student dropout setting. The LLM offers a promising classification algorithm, with a good trade-off between predictive accuracy and comprehensibility [20]. We test its performance, using a large, rich, real-life data set containing 10,554 students and 122 student variables from a global online learning provider, then benchmark that performance against eight predictive algorithms: the five most popular algorithms (see Table 2) and three algorithms that relate conceptually to LLM, namely, bagging or bootstrap aggregated decision trees (BAG), boosting trees (BOOST), and the logistic model tree (LMT). Second, we extend the visualization of LLM [20] to a learning analytics context to deliver better insights into students' actual learning drivers. Third, we investigate the impact of multiple student variables across several categories, including demographics and classroom characteristics, as well as cognitive, academic, and behavioral engagement. With these results, we suggest some ways to personalize student retention campaigns. The remainder of the paper is structured as follows. In Section 2, we outline the LLM, and its original and extended visualization. After we lay out the research procedure, data, experimental setting, and evaluation metrics, Section 4 details the predictive performance according to the benchmarking results and the extended LLM visualization. This article ends with a conclusion and directions for further research.

## 2. The logit leaf model and its visualization extension

The LLM is a two-step hybrid approach that automatically detects segments in the data by using the leaf nodes of a decision tree, then applies logistic regression to each segment [20]. Fig. 1 depicts its functioning in a student dropout setting. In the first step, a decision tree assigns different students to different segments. In the second step, a logistic regression is fit for every segment separately. The algorithm uses forward selection to choose the variables that enter the logistic regression for every segment separately, so it can detect segment-specific drivers. In an online learning subscription context, heterogeneity across student groups can cause different groups to be subject to multiple drivers of student dropout, so it provides a good test bed for the LLM [21]; we know of no studies that use the LLM in a learning analytics or online student dropout prediction context.

The LLM has several advantages over other algorithms. First, it is well suited to data sets that feature heterogeneity among students [20], because different students are assigned to different classifiers. Second, its built-in variable selection mechanism improves the predictive performance of the model. That is, it increases the stability of the model by reducing collinearity and preventing overfitting on noisy data [22]. As empirical research shows, models trained on well-selected variables tend to perform better than their counterparts trained on an extensive set that contains more noisy data [23]. In addition, reducing the number of variables results in more concise models that are easier to explain [24]. Third, the LLM is comprehensible and reveals specific drivers for each segment. Thus actionable insights can be derived and

**Table 2**
Summary of predictive algorithms in the online learning dropout literature.

| Predictive algorithm | Count |
|---|---|
| Logistic regression (LR) | 16 |
| Support vector machines (SVM) | 10 |
| Neural network (NN, DNN, RNN) | 8 |
| Decision tree (DT) | 6 |
| Random forest (RF) | 6 |
| Boosting tree (Boost) | 3 |
| Hidden Markov models (HMM) | 3 |
| Naïve Bayes (NB) | 3 |
| Bayesian network (BN) | 2 |
| Others | 6 |

segments can be managed according to their specific drivers. This comprehensible output also can be presented in a table with the decision rules that create the segments, together with the retained variables for each segment and their standardized Beta coefficients. These variables are grouped according to whether they are shared over different segments or unique for a particular segment. Table 3 offers an example of the original visualization for the LLM by De Caigny et al. [20], on fictitious data for illustrative purposes.

Although this visualization provides relevant student dropout information, it has several limitations. First, when the number of segments or retained variables increases, the table rapidly becomes unreadable, such that the main insights are lost. In this data, only three student segments are defined based on two variables (i.e. *Age* and *Average Grades*), but in reality there is more heterogeneity between students, resulting in an increased number of segments and number of DT variables. Therefore, the differences between segments become less interpretable, and less managerially actionable, in the original visualization when the number of variables and/or segments increases. Second, the original visualization of the LLM presents the drivers at a segment level. For example, dropout behavior in segment 3 depends on the participation rate (*Participation*), gender (*Gender_Female*), and homework completion percentage (*Homework Completion %*). This information would be sufficient, if we only considered heterogeneity *between* segments in our decision making process, but a student dropout setting often demands a more personalized approach, and heterogeneity *within* segments can be important too. Even if students belong to the same segment, their personal differences might be pertinent, but the original visualization cannot reveal these individualized insights.

We therefore propose a new LLM visualization that is better tailored to the student dropout case and resolves the limitations of the original visualization, as detailed in Table 4.

First, this revised version adds the dropout percentage per segment, due to its importance for comparing different segments. Second, it regroups the variables on the variable category level, reflecting the importance of each category in the new visualization. Third, the new visualization gets extended with a student-specific report to reveal individual drivers. The report is based on nomograms which have a long tradition in engineering and medicine to graphically represent the logistic regression calculations. A nomographic representation of a logistic regression is given as a series of straight individual arrows, where each arrow represents a variable. The individual arrows are all portrayed on a common linear scale to indicate the relative importance on the predicted dropout probability. The scale factors of the individual arrows are indicated by the length of the arrows a.k.a. the coefficients of the variables. The graphical evaluation of the logistic regression for a given student consists of locating the student values of the variables on the respective arrows, and determining corresponding point values on the common point scale. Nomograms help revealing and communicating the relative segment-specific effect of each variable, as well as the effect for an individual student [25].

In turn, the revised visualization offers three advantages for pedagogical decision makers. First, the First, the output remains readable, even if the number of variables increases, because we use the importance scores of the variable categories instead of raw variables. Second, differences between segments are indicated visually, because the added bar plots present relative importance scores for the variable categories. Third, the student-level reports using nomograms help users understand the relative effect of each variable, as well as its impact on each specific student. Our revised visualization is demonstrated on the real-life data and presented in Section 4.2.

## 3. Research procedure

### 3.1. Data

We conducted this research in collaboration with a global online learning provider, specializing in the field of data science. The data were extracted from a daily student interaction database, reflecting learning activities by 10,554 students over a period of 18 months. Students on this platform seek to learn Python, R, and SQL programming, as well as sharpen their computer and data science skills. They start their learning experience by subscribing to a monthly payment plan that grants them access to all courses, which they can start, complete, or quit in any order. The individual timelines for each student define the variety of their learning patterns and activities. Students' subscriptions renew automatically at the end of the subscription month; if students decide to stop their learning journey, they must do so before the end of the month. Therefore, we define the student dropout prediction problem as identifying students who will stop their learning journey and thus cancel the automatic payment plan for the next subscription month. We identify 5816 dropout cases, accounting for 55% of the total number of students.

To capture student learning activity, we rely on 122 student variables as independent variables in the student dropout prediction models, classified into demographics, classroom characteristics, and cognitive, academic, and behavioral engagement variables [18]. For each variable category, these variables might be available over two time windows: the student's full subscription duration or the last month. Table 5 summarizes the variable categories and number of variables per category.

### 3.2. Experimental setting

Data preprocessing is an essential phase that occurs directly before the predictive modeling step. It is time-consuming, but indispensable. It has proven to have an impact on prediction performance [26,27]. As is true for any predictive modeling setting, the well-known “garbage in, garbage out” rule applies to student dropout prediction. Literature proposes various strategies to deal with missing values; the deletion of students with missing values; the replacement of missing values with a constant or mean/mode value. This study follows the missing value replacement guidelines by [23]; a seminal paper in a closely-related domain to student dropout prediction. Missing values are treated differently depending on the percentage of missing values. For student variables with more than 5% missing values, we impute missing values. For continuous variables, we use 0, whereas the missing values of categorical variables are an additional category. For each variable for which we imputed missing values, we create a dummy variable to trace back imputation positions. For variables with less than 5% missing values, we remove these students, to reduce the impact of the imputation procedures. Furthermore, categorical variables are encoded by binary dummy variables. This method creates $c - 1$ dummy variables, where $c$ is the number of distinct categories of the student variable.

With our experimental design, we can compare the performance of the LLM against that of eight benchmark predictive models. Logistic regression (LR) is used as a benchmark given that it is one of the most

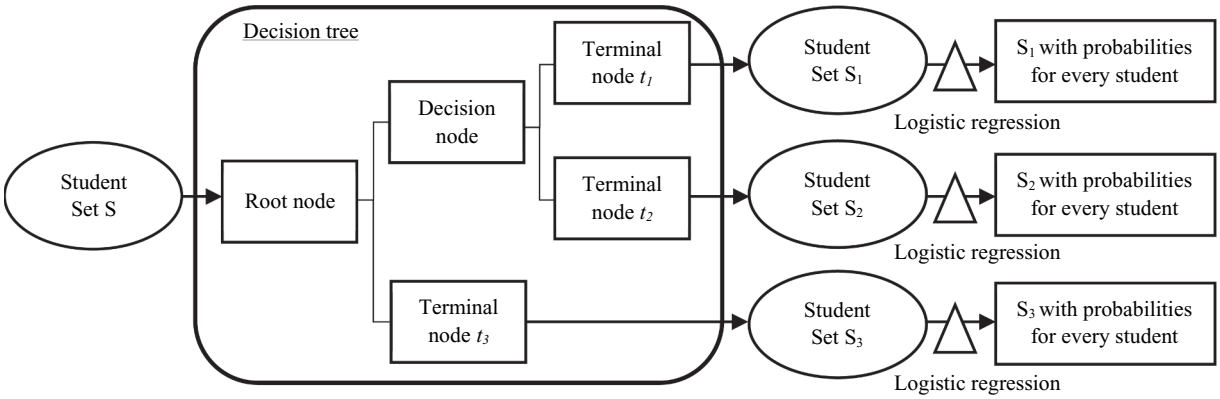



**Fig. 1.** Conceptual representation of the logit leaf model in a student dropout setting.

popular classification methods in learning analytics [28], and in online learning student dropout prediction in particular [29–31]. Moreover, a neural network (NN) has been used frequently in a student dropout context [32, 33] and can accurately identify students with potential difficulties. This study employs the feed-forward multilayer perceptron (MLP) NN [34,35]. Each independent (student) variable is a neuron in the input layer; the dependent variable that reflects student dropout behavior is represented in the output layer. We consider only one hidden layer as NN literature shows that one hidden layer is complex enough to be a universal function approximator [36]. A logistic activation function, used in between the hidden layer and the output layer, assigns a dropout probability to each student. The back-propagation algorithm is used as being the most popular NN optimizer [37]. In prior research [38], SVM has predicted students' performance with high accuracy, as well as student dropouts from massive open online courses (MOOCs) [39,40]. We employ the radial basis function (RBF) kernel as being the most popular kernel [41]. Further, decision trees offer advantages of simplicity and interpretability [42], and is often used as a baseline model in online learning contexts [29,39,43]. For our research, we consider the classification and regression tree (CART) and C4.5. To stabilize CART and C4.5, a final step prunes the tree according to the maximum performance objective criterion [44,45]. Next, ensembles of DTs are promising algorithms in field of learning analytics and student dropout prediction in particular [e.g. 39,46,47]. Bagging (BAG) is a statistical technique where an ensemble of DTs is built on bootstrap samples of the training set, also called bootstrap aggregation or bagging. This decreases the variance of the estimated prediction function [45,48]. Further, the combination of bootstrap sampling and a randomly picking of a predetermined number of input variables in an iterative way to create new training datasets that are afterwards modeled using CART DTs results in a powerful predictive algorithm called random forests (RF) that we also consider as one of our benchmarks [49]. Boosting (BOOST) aims to combine a set of weak DTs to create a powerful *committee* of DTs [50,51]. Instead of trying to fit the data hard like BAG and RF, BOOST learns data successively by repeatedly adjusting the weights of misclassified students in the training set. As the iterations proceed, students who are difficult to classify take on increasing influence, with higher weights assigned to them. Each successive classifier is thereby forced to concentrate on students that have been missed by previous ones in the sequence. Following BOOST variations are considered in this study, i.e. Adaboost, Logitboost, Gentleboost, and Robustboost [52,53]. A last benchmark algorithm is the logistic model tree (LMT). LMT is an alternative ensemble method and included as a benchmark given its resemblance with LLM. [54].

Table 6 summarizes the different meta-parameters for the different prediction models. The choice of prediction model parameters follows previous large-scale benchmarking studies [34]. We use a 5 × 2 cross-validation [23,34]. To optimize meta-parameters per fold, we split the training set further, such that we use two-thirds of the training set for meta-parameter tuning and the remaining one-third to evaluate and select the best combination of meta-parameters. This best combination then serves to train the prediction model on the full training set, scored on the test set of the fold.

### 3.3. Evaluation metrics

To compare the predictive performance of the LLM and eight benchmark models, we use the area under the receiver operating characteristic curve (AUROC or AUC) and the top decile lift (TDL) metrics [26,55]. The model performance comparison relies on a combined 5 × 2 cross-validation F-test [56].

To calculate the combined 5 × 2 cross-validation F-test, we approximate the F distribution with 10 and 5 degrees of freedom on the performance metrics AUC and TDL [56]. If $p < .05$, we can reject the null hypothesis that there is no significant difference in the performance metric values of the two prediction models or, in other words, that the two models differ significantly in their predictive performance, with a 95% confidence level.

**Table 3**
Illustration of the visualization logit leaf model.

| Seg | 1st step: decision tree | | | 2nd step: logistic regression | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Rule 1 | Rule 2 | # obs. | Shared Variables | | | | | Segment-Specific Variables |
| | | | | Intercept | Participation | | Homework Completion % | International | Grades High School |
| 1 | *Age* ≤ 20 | | 120 | 0.46 | −1.12 | | −0.28 | 0.63 | 0.23 |
| | | | | Intercept | Participation | Gender_Female | Homework Completion % | International | |
| 2 | *Age* > 20 | *Average Grades* ≤ 14 | 200 | 0.19 | −0.98 | −0.29 | −0.50 | −0.51 | |
| | | | | Intercept | Participation | Gender_Female | Homework Completion % | | |
| 3 | *Age* > 20 | *Average Grades* > 14 | 180 | 0.23 | −0.17 | −0.19 | −0.35 | | |

**Table 4**
Adjustments to the visualization logit leaf model.

| 1st Step: Decision tree | | | # obs. | 2nd Step: Logistic Regression | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Shared Variables | | | | | Segment Specific variables | |
| Seg | Rule 1 | Rule 2 | | Intercept | Participation | Gender_Female | Homework Completion % | International | Grades High School | |
| 1 | *Age* ≤ 20 | | 120 | 0.46 | -1.12 | | -0.28 | 0.63 | 0.23 | |
| 2 | *Age* > 20 | *Average Grades* ≤ 14 | 200 | 0.19 | -0.98 | -0.29 | -0.50 | -0.51 | | |
| 3 | *Age* > 20 | *Average Grades* > 14 | 180 | 0.23 | -0.17 | -0.19 | -0.35 | | | |

Adjustment 1: add dropout percentage per segment

Adjustment 2: replace the variables with the importance per variable category
Adjustment 3: visualize the impact of variables in a separate report using nomograms

## 4. Results

We present the predictive benchmark results on the real-life student dropout data. The second paragraph describes the extended visualization of the LLM model, which offers insights into student dropout prediction drivers.

### 4.1. Benchmarking

Table 7 presents the prediction evaluation of the LLM and eight benchmark models on the real-life subscription-based online learning data set. The AUC and TDL values confirm that it is possible to predict student dropout; all the prediction models perform better than random guessing, with AUC values greater than 0.5 and TDL values greater than 1. Including student demographics; classroom characteristics; and cognitive, academic, and behavioral engagement variables thus captures enough valuable information and enables all prediction models to differentiate dropout versus non-dropout students.

Table 8 shows the results of the combined 5 × 2 cross-validation F-test for the AUC and TDL between LLM and the eight benchmark models plus random guessing. The null hypothesis is that there is no significant difference in AUC (TDL) values between the LLM and the benchmark models. The results in Tables 7 and 8 affirm that LLM is significantly better than random guessing, and it is among the top performing algorithms. In combination with its excellent comprehensibility, this outcome identifies LLM as the algorithm that best balances predictive performance with comprehensibility for student variables. The results show that LLM is significantly better at detecting student dropout than DT, BAG, LMT, and NN; it performs equivalently well with LR, SVM, RF, and BOOST. Relative to the *single* predictive models LR and SVM, LLM accounts for heterogeneity in the customer base and thus can identify different drivers that explain student dropout for various student segments. In addition, LLM delivers a closed-form solution to the posterior dropout probabilities relative to SVM; its hyperparameter optimization process is slimmer and faster than SVM's. In contrast with the *ensemble* RF and BOOST models, LLM is easier and faster to train and score, and it outperforms them in the interpretability (and thus actionability) of the student drivers.

### 4.2. Interpretation and extended visualization

We interpret the results of the LLM using our proposed visualization. We start by calculating the relative variable category importance scores, to provide insights into what drives student dropout. The importance value of any single variable in the LLM can be calculated by taking the weighted sum of the Wald statistic of that variable in each segment. These importance values per variable are aggregated by variable category and divided by the total importance of all categories. Fig. 2 presents the overall, relative contribution of each variable category.

Academic engagement, including academic performance and interaction with the course content, contribute the most by far to the LLM's predictions. A student's active or passive learning mode (*Cognitive Engagement*) and profile (*Demographics*) contribute about equally, with total importance of around 10%. Course characteristics (*Classroom*) and other student online activities (*Behavioral Engagement*) have little importance (< 5%) in the student dropout prediction model. This analysis of variable category importance thus provides a good first impression of student dropout drivers.

To gain deeper insights on the segment level and devise a tailored student retention management plan, the LLM and its hybrid two-stage modeling approach also can provide insights about specific segments. This new visualization, shown in Figs. 3 and 4, improves on the original visualization in two ways.

First, it reports variable group importance values, instead of variable-specific Beta coefficients, so we can easily identify differences

**Table 5**
Variables framework to predict online learning student dropout.

| Variable categories | Variable time window | | Total number of variables |
|---|---|---|---|
| | Full subscription | Last month | |
| 1/ Demographics: variables related to student profile information (e.g., continent, avatar usage, number of subscription months) | 6 | – | 6 |
| 2/ Classroom characteristics: variables related to online classroom characteristics (e.g., course difficulty) | 8 | – | 8 |
| 3/ Cognitive engagement: variables related to student's learning strategy (e.g., active [time between two exercises, seeking for help/hint, studying in weekend], passive [seeking for solution]) | 19 | 5 | 24 |
| 4/ Academic engagement: variables related directly to student's online studying activities (e.g., course completion rate, number of exercises started/completed/uncompleted, number of attempts per exercise, total XP points collected) | 42 | 38 | 80 |
| 5/ Behavioral engagement: variables related to student's participation to other online activities (e.g., giving feedback by rating the exercise) | 2 | 2 | 4 |
| | 77 | 45 | 122 |

across the various segments. The output thus remains comprehensible, even with more variables and/or segments. As Fig. 3 clearly displays, four segments can be created on the basis of three decision rules. A closer inspection of the segments highlights segment 4, which contains students who maintain longer length of subscription (or *Number of previous subscriptions paid by student* $>$ 5), yet the academic engagement category is less important, and the demographics category is more important for predicting student dropout, relative to the other segments. This loyal segment accounts for the lowest dropout percentage, of 27%. Segment 3 also is interesting; it contains novice students (*Number of previous subscriptions paid by student* $\leq$ 5), who indicate a higher average number of attempts per exercise (*Average number of attempts per exercise during the subscription period* $>$ 2.57). It also represents the highest student dropout percentage, of 76%. Considering their many attempts, segment 3 appears to contain two types of students. On the one hand, students with a high learning motivation and dedication might have reached their learning goals and therefore stop subscribing. On the other hand, highly motivated students might struggle with learning difficulties, such that the exercises are too difficult, leading them to drop out as well. This explanation appears likely, considering that the academic engagement category is the most important for predicting student dropout in this segment.

Second, a report on the student level (Fig. 4) extends the LLM visualization. At the top, this report identifies the student and provides a brief summary of key characteristics of the segment to which the student belongs. It also provides a gauge chart that shows the predicted dropout probability according to the LLM. Significant variables appear on the left hand side; they are depicted with nomograms on the right. Critical criteria for building nomograms are as follows:

- All variables use the same scale to indicate their relative impact on the predicted probability, on a range from 1 to 100.
- The length of the arrow indicates the variable impact, and variables are sorted from most to least impact on student dropout behavior. The original range of the variable, or minimum and maximum values, is indicated above the arrow.
- Variables with a positive Beta coefficient have an arrow pointing to the right (i.e., higher value increases predicted dropout probability), whereas variables with a negative Beta coefficient have an arrow pointing to the left (i.e., lower value increases predicted probability of dropout).

**Table 7**
Predictive performance overview.

| Prediction model | AUC (SD) | TDL (SD) |
|---|---|---|
| LR | 0.845 (0.006) | 1.801 (0.020) |
| LLM | 0.839 (0.010) | 1.798 (0.022) |
| BOOST | 0.837 (0.005) | 1.786 (0.043) |
| SVM | 0.828 (0.005) | 1.750 (0.031) |
| RF | 0.826 (0.004) | 1.782 (0.017) |
| LMT | 0.801 (0.005) | 1.691 (0.042) |
| BAG | 0.780 (0.006) | 1.678 (0.079) |
| NN | 0.778 (0.007) | 1.651 (0.074) |
| DT | 0.728 (0.015) | 1.436 (0.092) |

**Table 6**
Prediction models and meta-parameters.

| Prediction Models | Number of model algorithms | Meta-parameter | Candidate settings |
|---|---|---|---|
| LR | 1 | n/a | n/a |
| DT | 46 | Decision Tree algorithm | CART, C4.5 |
| | | Min. leaf size | $n$*[0.01, 0.025, 0.05, 0.1, 0.25, 0.5] |
| | | Pruning (CART) | on/off |
| | | Confidence threshold for pruning (C4.5) | 0.01, 0.15, …, 0.30 |
| BAG | 9 | No. of bootstrap samples | 10, 20, …, 50, 100, 250, 500, 1000 |
| RF | 30 | No. of CART trees | 100, 250, 500, 750, 1000 |
| | | No. of randomly sampled variables | $\sqrt{m} * [0.1, 0.25, 0.5, 1, 2, 4]$ |
| BOOST | 48 | Boosting algorithm | AdaboostM1, Logitboost, Gentleboost, Robustboost |
| | | No. of boosting iterations | 10, 50, 100, 250, 500, 1000 |
| | | Learning rate (Gentleboost) | 0.1, 0.5, 1 |
| | | Max. margin (Robustboost) | 0.1 0.5*$p$(y = +1), $p$(y = +1) |
| LMT | 1 | n/a | |
| LLM | 36 | Confidence threshold for pruning | 0.01, 0.15, …, 0.30 |
| | | Min. leaf size | $n$*[0.01,0.025,0.05,0.1,0.25,0.5] |
| NN | 171 | No. of hidden nodes | 2, 3, …, 20 |
| | | Regularization penalty | $10^{(-4,\ -3.5,\ \ldots,\ 0)}$ |
| SVM | 300 | Regularization penalty | $2^{(-12,\ -13,\ \ldots,\ 12)}$ |
| | | Width of RBF kernel | $2^{(-12,\ -13,\ \ldots,\ -1)}$ |

**Table 8**
Combined 5 × 2 CV F-test, *p*-value on AUC and TDL values.

| | | LR | DT | BAG | RF | BOOST | LMT | SVM | NN | random guess |
|---|---|---|---|---|---|---|---|---|---|---|
| LLM | AUC | 0.25 | 0.00* | 0.00* | 0.19 | 0.62 | 0.00* | 0.43 | 0.00* | 0.00* |
| | TDL | 0.22 | 0.00* | 0.13 | 0.33 | 0.44 | 0.02* | 0.18 | 0.00* | 0.00* |

* $p < .05$.

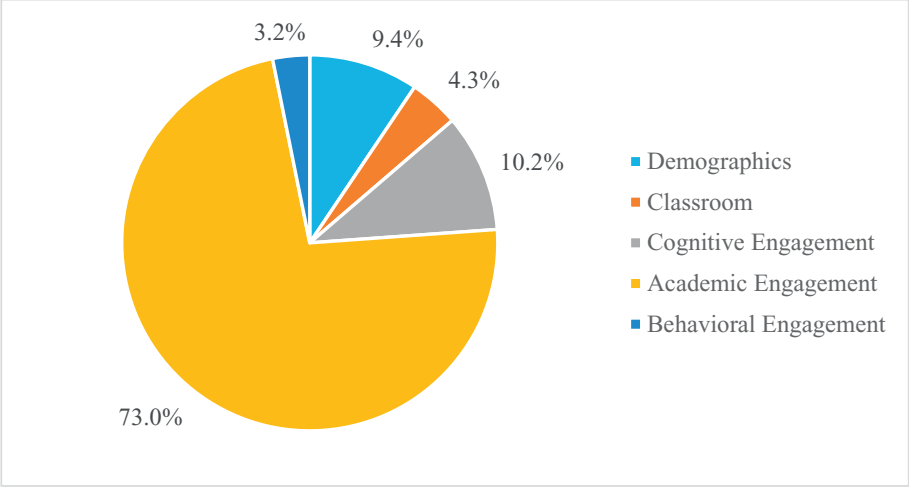



**Fig. 2.** Percentage of variable importance per variable group according to LLM model.

- The pointer ▲ indicates the variable value for the student.
- Variables are color coded, based on the associated variable category (cf. colors in Fig. 2)

This report on the student level provides a tool for analyzing dropout drivers for an individual student, which was lacking in the original LLM visualization. Fig. 4 shows the report for student 3697, who belongs to segment 1—characterized by few subscriptions, an average number of attempts per exercise, and average course difficulty. The visualization indicates that the student's predicted dropout probability is 76%, obtained from the LLM model. At a glance, the visualization offers deeper insights into the importance of the student dropout drivers; the color coding makes immediately clear which variable categories are most impactful for this segment. For example, in Fig. 4, most impactful variables for segment 1 come from the academic engagement category (yellow coded). The variables also are clearly ranked from highest to lowest impact on the student dropout behavior of the *segment* (see also the length of the arrow in the nomogram), and the pointer ▲ indicates, for this *student*, the degree to which the variable influences dropout behavior. The more the pointer deviates from the middle, the higher the variable's impact is. For student 3697, the specific variables that contribute to elevated dropout risk mainly result from the student's lack of clear understanding of the learning objectives and intentions. The student starts multiple chapters at the same time (*Average number of chapters started at the same time*), without having completed any courses at the start of the latest subscription (*Count of number of completed courses at start of latest subscription*). From a decision making perspective, this is important information. This behavior indicates that the student has difficulties setting learning goals and as a result is aimlessly floating around the online learning environment. As the model predicts that this type of behavior has a high probability of ending in drop-out, intervention by the system should attempt to get the student back on track. This could be done by suggesting the student several learning paths or a logical sequence of courses, and thus consequently encouraging or obliging students to finish a module before starting a new one. The nomograms in Fig. 4 indicate which interventions are relevant for the particular student. The triangles for the right arrows should be pulled as much as possible to 0 points, while the inverse is true for the triangles on the left arrows. In the case of student 3697, the valuable suggestion is to focus student's attention on a specific single course, ignoring the other started courses and chapters. This could be implemented in the online learning

| Segment Definition | | | | | | Variable Category Importance | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Seg | Rule 1 | Rule 2 | Rule 3 | Number of Observations | Percentage Dropout | Demographics | Classroom | Cognitive Engagement | Academic Engagement | Behavioral Engagement |
| | *Number of previous subscriptions paid by student (V1)* | *Average number of attempts per exercise by student during the subscription period (V2)* | *Average difficulty level of all courses taken by the student during the subscription (V3)* | | | | | | | |
| 1 | $V1 \leq 5$ | $V2 \leq 2.57$ | $V3 \leq 1.69$ | 1,241 | 64% | 12% | 2% | 18% | 68% | 0% |
| 2 | $V1 \leq 5$ | $V2 \leq 2.57$ | $V3 > 1.69$ | 1,268 | 45% | 7% | 3% | 11% | 77% | 2% |
| 3 | $V1 \leq 5$ | $V2 > 2.57$ | | 1,264 | 76% | 1% | 7% | 3% | 85% | 4% |
| 4 | $V$ | | | 1,000 | 27% | 30% | 4% | 11% | 46% | 9% |

**Fig. 3.** Percentage of variable importance per group of each segment of LLM model.

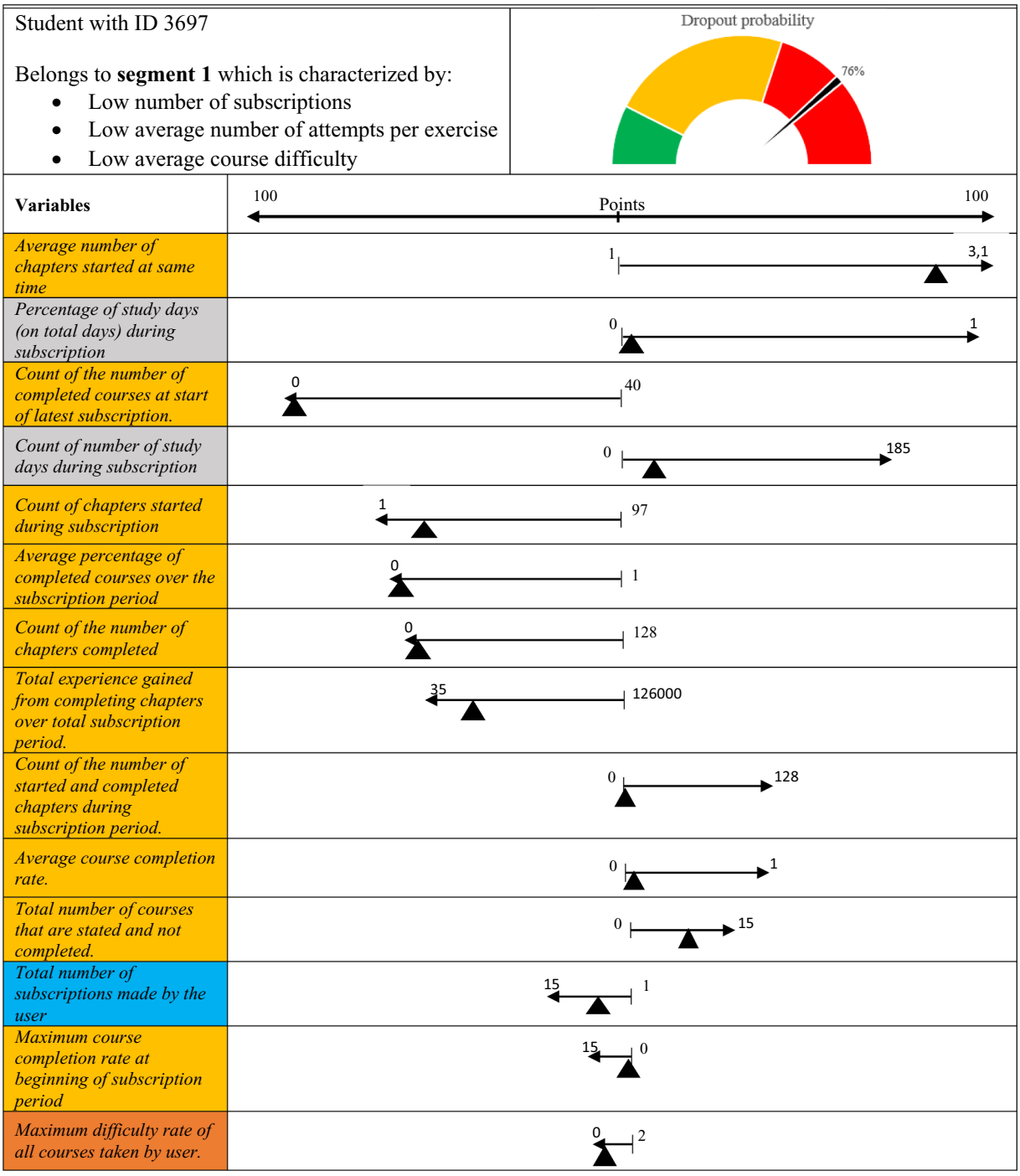



**Fig. 4.** Visualization of student profile using the results of the LLM model.

platform by building in constraints to make sure that students finish a module before starting new ones. Also, the nomograms indicate that the student might benefit from notifications to motivate him/her to get back to the study material after some period of inactivity. In sum, educational decision makers could study the student profiles and make specific recommendations to student, or, for often observed profiles, some standardized suggestions could be integrated in the learning management system.

## 5. Conclusions

Learning analytics has gained considerable attention in educational and data science communities, and student dropout prediction is a popular application domain. This study contributes to extant literature and practice by offering insights for online learning providers to help them better manage student dropout risks. The contributions are threefold. First, the results of our study confirm that it is possible to predict student dropout accurately in a subscription-based online learning context using five types of variables: demographics; classroom characteristics; and cognitive, academic, and behavioral forms of engagement. The LLM and all benchmark models outperform random guessing, with strong prediction performance in terms of both AUC and TDL. Second, the benchmarking results confirm that the LLM offers the best trade-off between prediction performance and interpretability. That is, it outperforms all benchmark algorithms except LR, SVM, RF, and BOOST, but the latter algorithms offer less insight into the

heterogeneity of the student base. Third, the new LLM visualizations (Figs. 3 and 4) provide quick but deep insights into student dropout behavior. Online learning providers can zoom in on segment- and student-level specific drivers that will help them design relevant student dropout prevention actions.

This study thus delivers significant contributions to decision support systems literature; we also note several paths for continued research. First, our conclusions reflect an online subscription-based learning context. Additional research is needed to validate and test the beneficial effect of LLM in other learning analytics contexts. Second, this study proposes a new visualization tool for LLM based on nomograms to profile students (*descriptive analytics*) and focuses on predicting student dropouts (*predictive analytics*). However, prescriptive analytics, a.k.a. uplift or net-effect modeling, is getting more attention in the decision support systems literature [e.g. 57]. Unlike our predictive modeling paper that predicts students who will dropout, uplift modeling focusses on identifying those students who will dropout *and* could be retained as a result of the student retention action. In order words, an uplift model has the purpose to identify these students for whom a student retention action has the biggest impact. Therefore, a valuable path for future research is to extent the findings of this work by setting up an experimental design to investigate the impact of various student retention actions on future dropout behaviors. Third, the extended visualization of the LLM makes use of nomograms to give insights into the student-specific drivers of online dropout. However, various alternatives to nomograms have been proposed to boost the local interpretability of a classifiers, i.e. the impact of the variable importance on the student level. Future research could thus further extent the visual interpretation of LLM by plugging recently-developed methods like **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations (LIME) [58] or the **SH**apley **A**dditive Ex**P**lanation (SHAP) [59].

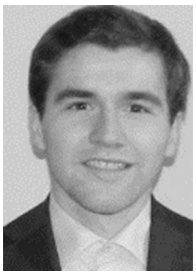
## References


[1] I.E. Allen, J. Seaman, R. Poulin, T.T. Straut, Online report card: tracking online education in the United\nStates, Retrieved March. 23 (2016), p. 2016.
[2] M. Reinig, The Theory and Practice of Online Learning, University of Washington Press, 2010.
[3] D.E. Simmons, The Forum Report: E-Learning Adoption Rates and Barriers, ASTD E-Learning Handb., 2002, pp. 19–23.
[4] A.W. Bates, Restructuring the University for Technological Change, Carnegie Found. Adv. Teach. What Kind Univ., 1997, pp. 207–228.
[5] Y. Lee, J. Choi, A review of online course dropout research: implications for practice and future research, Educ. Technol. Res. Dev. 59 (2011) 593–618.
[6] J.H. Park, H.J. Choi, Factors influencing adult learners' decision to drop out or persist in online learning, Educ. Technol. Soc. 12 (2009) 207–217.
[7] K.S. Hone, G.R. El Said, Exploring the factors affecting MOOC retention: a survey study, Comput. Educ. 98 (2016) 157–168.
[8] Y. Levy, Comparing dropouts and persistence in e-learning courses, Comput. Educ. 48 (2007) 185–204.
[9] J. Meister, Pillars of e-Learning Success, New York Corp. Univ. Exch., 2002.
[10] C. Parr, Mooc completion rates 'below 7%', Times High. Educ, 2013, pp. 7–9.
[11] B. Poellhuber, M. Chomienne, T. Karsenti, The effect of peer collaboration and collaborative learning on self-efficacy and persistence in a learner-paced continuous intake model, J. Distance Educ. 22 (2008) 41–62.
[12] S.Y. Liu, J. Gomez, C.J. Yen, Community college online course retention and final grade: predictability of social presence, J. Interact. Online Learn. 8 (2009) 165–182.
[13] P.A. Willging, S.D. Johnson, Factors that influence students' decision to dropout of online courses, J. Asynchronous Learn. Networks. 13 (2009) 115–127.
[14] A. Ghazal, T. Rabl, M. Hu, F. Raab, M. Poess, A. Crolotte, H.A. Jacobsen, BigBench: towards an industry standard benchmark for big data analytics, Proc. ACM SIGMOD Int. Conf. Manag. Data. 36 (2013) 1197–1208.
[15] K. Coussement, D.F. Benoit, D. Van den Poel, Improved marketing decision making in a customer churn prediction context using generalized additive models, Expert Syst. Appl. 37 (2010) 2132–2143.
[16] A. Gandomi, M. Haider, Beyond the hype: big data concepts, methods, and analytics, Int. J. Inf. Manag. 35 (2015) 137–144.
[17] S. Ryu, Book Review: Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie or Die, Wiley, New Jersey, 2013.
[18] S. Joksimović, O. Poquet, V. Kovanović, N. Dowell, C. Mills, D. Gašević, S. Dawson, A.C. Graesser, C. Brooks, How do we model learning at scale? A systematic review of research on MOOCs, Rev. Educ. Res. 88 (2017) 43–86.
[19] P. Long, G. Siemens, Penetrating the fog: analytics in learning and education, Educ. Rev. 46 (2011) 30–32.
[20] A. De Caigny, K. Coussement, K.W. De Bock, A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, Eur. J. Oper. Res. 269 (2018) 760–772.
[21] C. Masci, G. Johnes, T. Agasisti, Student and school performance across countries: a machine learning approach, Eur. J. Oper. Res. 269 (2018) 1072–1085.
[22] G. James, D. Witten, T. Hastie, R. Tibishirani, An Introduction to Statistical Learning with Applications in R (Older Version), Springer Publishing Company, Incorporated, 2013.
[23] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens, New insights into churn prediction in the telecommunication sector: a profit driven data mining approach, Eur. J. Oper. Res. 218 (2012) 211–229.
[24] S. Lessmann, S. Voß, A reference model for customer-centric data mining with support vector machines, Eur. J. Oper. Res. 199 (2009) 520–530.
[25] V. Van Belle, B. Van Calster, Visualizing risk prediction models, PLoS One 10 (2015) e0132614.
[26] K. Coussement, S. Lessmann, G. Verstraeten, A comparative analysis of data preparation algorithms for customer churn prediction: a case study in the telecommunication industry, Decis. Support. Syst. 95 (2017) 27–36.
[27] D. Pyle, S. Editor, D.D. Cerra, Data Preparation for Data Mining, Morgan Kaufmann Publishers, San Francisco, 1999.
[28] R. Barber, M. Sharkey, Course correction: using analytics to predict course success, ACM Int. Conf. Proceeding Ser, ACM, New York, NY, USA, 2012, pp. 259–262.
[29] S.B. Kotsiantis, C.J. Pierrakeas, P.E. Pintelas, Preventing student dropout in distance learning using machine learning techniques, Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci.), 2003, pp. 267–274.
[30] S. Jiang, A.E. Williams, K. Schenke, M. Warschauer, D.O. Dowd, Predicting MOOC performance with week 1 behavior, Proc. 7th Int. Conf. Educ. Data Min, 2014, pp. 273–275.
[31] K.R. Koedinger, E.A. McLaughlin, J. Kim, J.Z. Jia, N.L. Bier, Learning is not a spectator sport: doing is better than watching for learning from a MOOC, L@S 2015 - 2nd ACM Conf. Learn. Scale, 2015, pp. 111–120.
[32] D. Delen, A comparative analysis of machine learning techniques for student retention management, Decis. Support. Syst. 49 (2010) 498–506.
[33] A.S. Hoffait, M. Schyns, Early detection of university students with potential difficulties, Decis. Support. Syst. 101 (2017) 1–11.
[34] S. Lessmann, B. Baesens, H.V. Seow, L.C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research, Eur. J. Oper. Res. 247 (2015) 124–136.
[35] D. Ruppert, The elements of statistical learning: data mining, inference, and prediction, J. Am. Stat. Assoc. 99 (2004) 567.
[36] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Inc., New York, NY, USA, 1995.
[37] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representation by error propagation. Parallel distributed processing: foundations, vol. 1, (1986).
[38] V.L. Miguéis, A. Freitas, P.J.V. Garcia, A. Silva, Early segmentation of students according to their academic performance: A predictive modelling approach, Decis. Support Syst. 115 (2018) 36–51.
[39] W. Wang, H. Yu, C. Miao, Deep model for dropout prediction in MOOCs, ACM Int. Conf. Proceeding Ser, 2017, pp. 26–32.
[40] J. Liang, C. Li, L. Zheng, Machine learning application in MOOCs: Dropout prediction, ICCSE 2016 - 11th Int. Conf. Comput. Sci. Educ, 2016, pp. 52–57.
[41] B. Schölkopf, K.K. Sung, C.J.C. Burges, F. Girosi, P. Niyogi, T. Poggio, V. Vapnik, Comparing support vector machines with gaussian kernels to radial basis function classifiers, IEEE Trans. Signal Process. 45 (1997) 2758–2765.
[42] K. Coussement, K.W. De Bock, Customer churn prediction in the online gambling industry: the beneficial effect of ensemble learning, J. Bus. Res. 66 (2013) 1629–1636.
[43] M. Tan, P. Shao, Prediction of student dropout in E-learning program through the use of machine learning method, Int. J. Emerg. Technol. Learn. 10 (2015) 11–17.
[44] D. Steinberg, CART: classification and regression trees, Top Ten Algorithms Data Min. 9 (2009) 179.
[45] E.R. Ziegel, The Elements of Statistical Learning, Springer, New York, 2003.
[46] R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn, N. Radi, Machine learning approaches to predict learning outcomes in Massive open online courses, Proc. Int. Jt. Conf. Neural Netw, 2017, pp. 713–720.
[47] S. Nagrecha, J.Z. Dillon, N.V. Chawla, MOOC dropout prediction: Lessons learned from making pipelines interpretable, 26th Int. World Wide Web Conf. 2017, WWW 2017 Companion, 2019, pp. 351–359.
[48] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[49] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[50] M. Kearns, Thoughts on hypothesis boosting, Unpubl. Manuscr. 45 (1988) 105.
[51] M. Kearns, L. Valiant, Cryptographic limitations on learning Boolean formulae and finite automata, J. ACM 41 (1994) 67–95.
[52] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 904 (1995) 23–37.
[53] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Ann. Stat. 28 (2000) 337–407.
[54] N. Landwehr, M. Hall, E. Frank, Logistic model trees, Mach. Learn. 59 (2005) 161–205.
[55] P. Hainaut, B. Vozar, S. Rinaldi, E. Riboli, E. Caboux, The European prospective investigation into cancer and nutrition biobank, Methods Mol. Biol. 675 (2011) 179–191.
[56] E. Alpaydin, Combined 5 x 2 cv F test for comparing supervised classification learning algorithms, Neural Comput. 11 (1999) 1885–1892.
[57] S. Debaere, F. Devriendt, J. Brunneder, W. Verbeke, T. De Ruyck, K. Coussement, Reducing inferior member community participation using uplift modeling: evidence

from a field experiment, Decis. Support Syst. 123 (2019).
[58] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 13–17-Augu, 2016, pp. 1135–1144.
[59] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst, 2017, pp. 4766–4775.
[60] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, V. Loumos, Dropout prediction in e-learning courses through the combination of machine learning techniques, Comput. Educ. 53 (2009) 950–965.
[61] G. Balakrishnan, D. Coetzee, Predicting student retention in massive open online courses using hidden markov models, Electr. Eng. Comput. Sci. Univ. Calif. Berkeley, 2013, pp. 1–15.
[62] M. Sharkey, R. Sanders, A process for predicting MOOC attrition, Proc. EMNLP 2014 Work. Anal. Large Scale Soc. Interact. MOOCs, 2015, pp. 50–54.
[63] B. Amnueypornsakul, S. Bhat, P. Chinprutthiwong, Predicting attrition along the way: the UIUC model, Proc. EMNLP 2014 Work. Anal. Large Scale Soc. Interact. MOOCs, 2015, pp. 55–59.
[64] M. Kloft, F. Stiehler, Z. Zheng, N. Pinkwart, Predicting MOOC dropout over weeks using machine learning methods, Proc. EMNLP 2014 Work. Anal. Large Scale Soc. Interact. MOOCs, 2015, pp. 60–65.
[65] J. Whitehill, J.J. Williams, G. Lopez, C.A. Coleman, J. Reich, Beyond Prediction: First Steps Toward Automatic Intervention in MOOC Student Stopout, SSRN Electron. J, (2015).
[66] J. He, J. Baileyt, B.I.P. Rubinstein, R. Zhang, Identifying at-risk students in massive open online courses, Proc. Natl. Conf. Artif. Intell, 2015, pp. 1749–1755.
[67] D.S. Chaplot, E. Rhim, J. Kim, Predicting student attrition in MOOCs using sentiment analysis and neural networks, CEUR Workshop Proc, 2015, pp. 7–12.
[68] S. Boyer, K. Veeramachaneni, Transfer learning for predictive models in massive open online courses, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2015, pp. 54–63.
[69] M. Fei, D.Y. Yeung, Temporal models for predicting student dropout in massive open online courses, Proc. - 15th IEEE Int. Conf. Data Min. Work. ICDMW, 2016 2015, pp. 256–263.
[70] W. Xing, X. Chen, J. Stein, M. Marcinkowski, Temporal predication of dropouts in MOOCs: reaching the low hanging fruit through stacking generalization, Comput. Human Behav. 58 (2016) 119–129.
[71] J. Qiu, J. Tang, T.X. Liu, J. Gong, C. Zhang, Q. Zhang, Y. Xue, Modeling and predicting learning behavior in MOOCs, WSDM 2016 - Proc. 9th ACM Int. Conf. Web Search Data Min, 2016, pp. 93–102.
[72] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, H. Gehlbach, Forecasting student achievement in MOOCs with natural language processing, ACM Int. Conf. Proceeding Ser, 2016, pp. 383–387.
[73] C. Burgos, M.L. Campanario, D. de la Peña, J.A. Lara, D. Lizcano, M.A. Martínez, Data mining for modeling students' performance: a tutoring action plan to prevent academic dropout, Comput. Electr. Eng. 66 (2018) 541–556.

**Prof. Kristof Coussement**, Ph.D., is Professor of Business Analytics, Director of the IESEG Center for Marketing Analytics, and Academic Director of the MSc in Big Data Analytics for Business at IESEG School of Management (LEM-CNRS). He is publishing in international peer reviewed journals such as Decision Support Systems, Information & Management, International Journal of Information Management, Information Sciences, Data Mining and Knowledge Discovery, International Journal of Forecasting, Knowledge-based Systems, European Journal of Operational Research, Journal of Product Innovation Management, Journal of Business Research, Research Policy, European Journal of Marketing, Computational Statistics & Data Analysis, Expert Systems with Applications, among others. Moreover, his works have been presented at various conferences around the world. His main research interests are all data science aspects applied in business.

**Minh Phan** is a Ph.D. candidate at IESEG School of Management (LEM-CNRS 9221) - Catholic University of Lille (Lille, France) (EQUIS, AACSB, AMBA). His research interests focus on machine learning applications for learning analytics. Currently, Minh is working on student dropout prediction using the educational data of his home institution and other online learning providers. Before starting his PhD, Minh worked for two years in Financial and Consulting industry in Vietnam (KPMG) and obtained the Master of Science degree in Big Data Analytics at IESEG School of Management.



**Dr. Arno De Caigny**, Ph.D., is Assistant Professor of Business Analytics at IESEG School of Management (LEM-CNRS 9221) - Catholic University of Lille (Lille, France) (EQUIS, AACSB, AMBA). He worked together with Crédit Agricole Nord de France in the context of the Big Data & Digital Banking chair. Before his PhD, Arno worked for one year as an analytical consultant at Deloitte in Belgium. Arno's research interests involve the improvement of customer scoring models using (big) data. He has published in international peer-reviewed journals such as European Journal of Operational Research, Decision Support Systems and International Journal of Forecasting.

Dr. **Dries F. Benoit** (Ph.D.) is Associate Professor in Data Analytics at Ghent University and teaches Bayesian statistics, statistical modeling & datamining and pricing & revenue management to the students in business engineering. He also is visiting professor at Université de Namur (Namur, Belgium). Dries Benoit specializes in Bayesian statistics: an alternative (to frequentist/classical statistics) paradigm for doing inference, prediction and model selection. He works on both methodological as well as applied problems, where most applications are in the field of business administration and management (marketing, finance, operations research). As a data scientist, he often works together with researchers from other fields such as medicine, energy, education, etc.

**Dr. Annelies Raes** holds a Ph.D. in Educational Technology by Ghent University and is currently working as Postdoctoral Researcher at the Centre for Instructional Psychology and Technology (CIP&T) at the University of Leuven (KU Leuven), campus Kulak in Kortrijk, Belgium. Annelies Raes is also co-Principal Investigator within IMEC's Smart Education Program (https://www.imec-int.com/en/articles/smart-education). Her main fields of interest are new innovative education models as active learning and problem-based collaborative learning and how this can be supported by emergent technologies. More specifically, her research focuses on the effective and efficient use of educational technologies. It encompasses the development and testing of smart technologies (sensors, algorithms, adaptive learning platforms, etc.) that facilitate interaction and collaboration in the learning process and lay the foundation of tailor-made learning solutions.