USING MACHINE LEARNING TO PREDICT MATHEMATICAL PERFORMANCE

Martin Fröhlich, Stefan Krauss, and Sven Hilbert
University of Regensburg, Germany
martin.froehlich@mathematik.uni-regensburg.de

In empirical educational research, it is of great interest to predict student performance. In contrast to other disciplines, however, machine learning methods for identifying promising predictors are not yet widely used. We will use a machine learning approach to study the effect of learning strategies and cooperative behaviors of German mathematics students with respect to exam grades. A small outlook will be given on how machine learning methods can be integrated into the education of young researchers in empirical educational research.

INTRODUCTION

In general, educational research is particularly engaged in identifying and evaluating important predictors of student achievement. Because the success of educational systems is of utmost relevance for societies (e.g., Hanushek & Woessmann, 2012), a flurry of empirical research consequently investigates achievement measures such as mathematical, scientific, or reading competencies as dependent variables (Hilbert et al., 2021).

In order to empirically identify important predictors for student learning gains, in principle, several experimental designs can be implemented. One option, for instance, is to systematically manipulate specific independent variables (IV) (e.g., teaching methods or learning strategies), such as in a pre-post design, and to subsequently establish the differential effects of this variation on student performance measures. Yet, following such designs, only a limited number of possible factors can be implemented and analyzed. (A famous meta-analysis regarding the results of such controlled experiments with respect to explaining achievement gains is Hattie & Yates, 2013.)

Another possibility to detect predictors of students' learning success is to provide data of so-called large-scale assessments, for example the Programme for International Student Assessment (PISA; https://www.oecd.org/pisa/) or the Trends in International Mathematics and Science Study (TIMSS; https://nces.ed.gov/timss/). These studies often implement a considerable quantity of scales and items, for instance on teacher competencies, instructional quality, and student learning gains, and thus allow for various correlational analyses (e.g., the COACTIV study; Kunter et al., 2013).

Note that after data collection in both designs mentioned so far (i.e., experimental pre-post designs versus correlational large-scale designs), data typically are analyzed by means of (linear) regression methods, which can establish the relationship between possible predictors and achievement measures in terms of regression coefficients. The main difference between the two approaches is that analyses after experimental designs often rely on a univariate model equation (typically a regression or ANOVA model) where IVs and some covariates are modelled as predictors and student achievement as dependent variable (DV). In contrast, data of large-scale studies are oftentimes analyzed via path or structural equation models (SEM), which can include a larger number of (latent) constructs. For instance, in path models and SEMs, mutual dependencies between IVs and DVs (exogenous and endogenous variables) are modeled by implementing *several regression models simultaneously*.

However, even in these more sophisticated approaches, the number of variables that can be analyzed simultaneously is limited. For instance, higher-dimensional SEM often run into convergence problems due to multi-collinearity (see, Kline, 2016). Therefore, even after applying experimental or correlational designs with "only" 30 variables (or scales, respectively, which is still far away from "large scale" data), it is almost impossible to implement all of the data in one common model. Therefore, in empirical educational research, usually a subset of the measured variables is selected for statistical analysis, whereas other variables do not find their way into the models. Although this variable selection is mostly justified in some theoretical way, it remains arbitrarily and untransparent, at least to a certain degree—especially if the non-included variables seem to be equally promising on a theoretical level. (It is hardly ever mentioned how many and which SEMs or regression models failed before the published one was "detected.")

In the present paper, we propose machine learning (ML) as a data-driven alternative for predicting student learning gains. ML provides a powerful, novel framework that can overcome many

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

shortcomings of classical methods of statistical inference, especially regarding restrictions with respect to the number of variables that can be included. Since ML is not yet widely implemented in university studies of young scientists in the educational sciences, in the final part we additionally recommend implementing such methods into the training of doctoral students, especially in pedagogy and the education of school disciplines.

In order to illustrate both the power and flexibility of ML approaches, we will demonstrate all of the following deliberations and analyses using an exemplary data set. A sample of $N = 90$ students of mathematics in a German university (University of Regensburg) was examined with respect to altogether $n = 154$ learning strategies and cooperative behaviors with fellow students (each implemented as single item measures, e.g., Table 2) to predict their final grade in the closing examination (see Table 1 for an overview of the design).

In the next section, however, we first give a short overview of ML, including its philosophy and its advantages, but also some typical critiques.

MACHINE LEARNING

In parallel with the growing computational power in the last decade, ML has turned into a strong competitor for inferential statistics when it comes to predicting dependent variables. In some fields (e.g., information and health sciences, cf. Hilbert et al., 2021), application of ML already constitutes the standard, whereas in educational sciences, there has not yet been such a coming shift in analytical paradigms. Although classical statistical regression methods, as mentioned, can only represent (generalized) linear relationships between a restricted number of variables, ML methods typically include more high-dimensional relationships without limiting the number of variables. A disadvantage of high-dimensional, non-linear models, which has evoked criticism regarding their usefulness, is the lack of parameters that allow for straightforward interpretation of the effect of individual features (the ML term for covariates) of the model. Yet, lately, the field of interpretable ML (iML, see, Molnar, 2020) has been flourishing and an approach has been derived to quantify the influence of individual model components even in complex, high-dimensional, non-linear situations.

Further advantages of ML are the easy detection and representation of non-linear relationships (such as depicted in Figure 2). In addition, the incorporation of elaborate (nested) resampling techniques in combination with prediction-centered model evaluation, predestines ML approaches for the development of thoroughly validated models that are replicable with novel data (Hilbert et al., 2021).

Therefore, more than just providing new analytical techniques, ML can help educational researchers change the modelling culture towards a stronger focus on robust models with reliable predictions instead of (over-)fitting complex but inflexible models to the dataset at hand (Yarkoni & Westfall, 2017).

DESIGN

In the lecture, "Linear Algebra II," at the faculty of mathematics at the University of Regensburg (Bavaria), a pre-post design including three measurement points was implemented (Table 1). However, this was not an experimental design (because no IV was systematically varied) but rather a correlational study aiming to investigate the effects of possible factors that could be relevant for final success on the exam. These factors were identified beforehand via both a literature review (e.g., Liebendörfer et al., 2021) and an informal Delphi-Study within the faculty, in which members could propose specific factors that in their opinion are crucial for learning success regarding mathematics. See Table 2 for five example items out of the 154 strategies and behaviors for which data were collected.

Table 1. Overview on the design of the study

Measurement Points		
1	2	3
(2nd week of semester)	(10th week of semester)	(15th week of semester)
Pre-test: Linear algebra	Learning strategies and cooperative behavior *during the semester*	Learning strategies and cooperative behavior *specific for exam preparation* Post-test: Linear algebra (final grade)

Table 2. Sample items: Learning strategies and cooperation with fellow students (self-reports)

Items (5 out of 154)	[scale: 1 = disagree; 2 = somewhat disagree; 3 = somewhat agree; 4 = agree]
s21d2_lserkla	I usually can explain the content of definitions and theorems in my own words.
s21d3_awbewei	For the exam preparation I have memorized the following: Proofs.
s21d2_lsauswd	I usually learn definitions by heart.
s21d2_lsaubsp	I usually check statements with examples.
s21d3_stunden	How many hours in total (do you estimate) did you study for the exam?

RESULTS

Implementing classical regression models would require the selection of a limited number of variables (and therefore a specific variable set) as predictors in the model. Note that based on 154 strategies and behaviors (all assessed as single item measures), almost more than half a million different single regression models would be possible if one would (arbitrarily) fix the number of implemented items to, say, four (with one of them being the pre-tests). In contrast, ML methods allow for implementing *all strategies simultaneously*. As a first step, a feature importance plot can display the variables ordered according to their impact on the final grade (Figure 1).

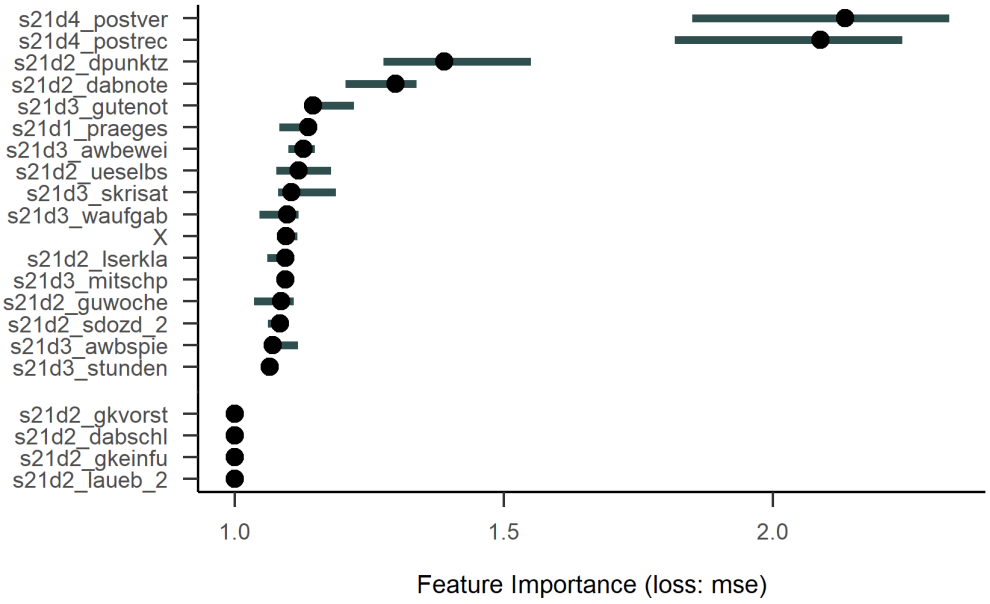
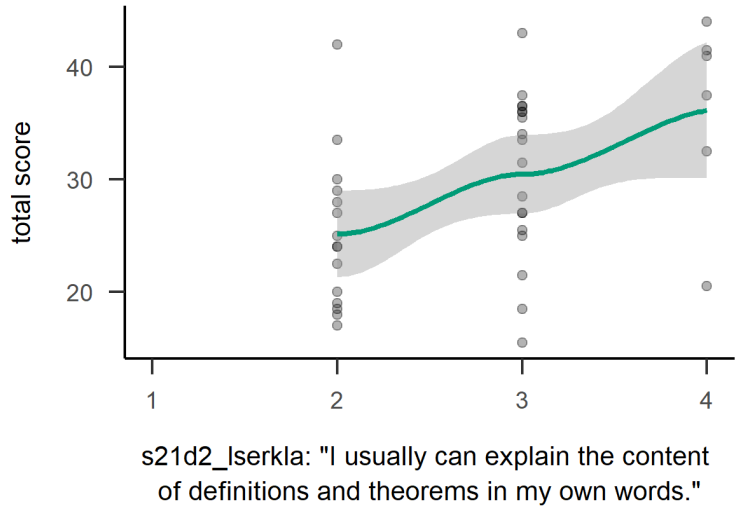
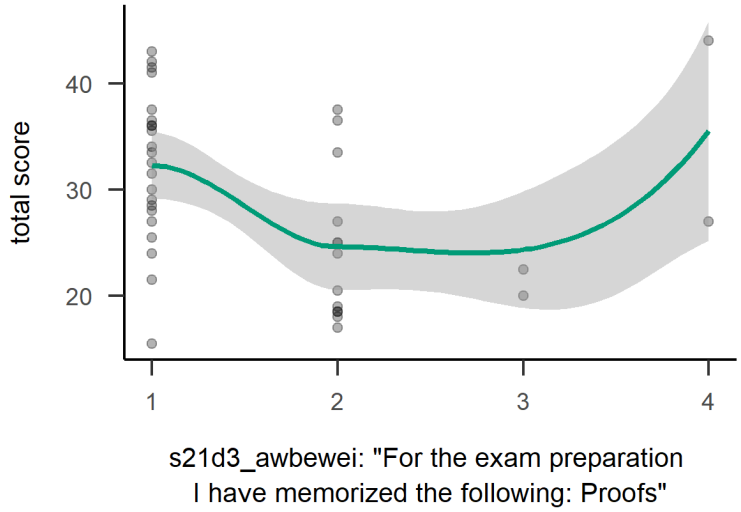


Figure 1. A feature importance plot with the items ordered according to their impact on the final grade. The 17 most impactful and 4 least impactful items are presented. The Mean Squared Error (MSE) is used as loss function.

In a next step, for instance, the items might be separated into context factors (that cannot be influenced by the students and/or teachers) and de facto learning strategies and cooperative behaviors (that students could judge from 1 = disagree until 4 = agree). The reason is that only results on the latter ones (the self-reported strategies and behaviors) allow for specific didactical hints that can be given to students and/or teachers. The most predictive items of the latter category then can be displayed as scatterplots with a local regression spline (see Figure 2), for which, however, ML is not yet necessarily required.



(a)



(b)

Figure 2. Two scatterplots, each with a local regression spline plot concerning the two most predictively valid learning strategies

First, the non-linearity of the relationship in the Figure 2(b) plot is intuitively understandable and very informative. This alone provides a more sophisticated insight into the relationship of variables (as compared to classical linear regression).

Second, based on the data collected so far, the (self-reported) 'ability to explain the content of definitions and theorems of the lecture in own words' (Figure 2(a)) was the best predictor of the final grade, when simultaneously all other strategies entered in the model. Interestingly, if a student said that he or she memorized the proofs of theorems by heart (Figure 2(b)), this had a *negative* effect on the final grade (especially when the outlier at category 4 would be neglected).

In addition, it easily becomes clear that a larger sample would be needed (in the preprocessing phase the sample had to be reduced to $N = 39$), because in category 4 only two students are depicted. Here, just 2–3 more students could make a great difference with respect to the shape of the curve (in summer 2022 the sample is extended by a replication study).

CONCLUSION

The ML algorithm could readily detect items that had a large impact on the total score of the final grade. It was very easy in the follow-up to look more closely at the most promising strategies and behaviors. In addition, not only could ML methods be used to easily identify non-linear relationships

but also to represent such relations in a very intuitive way. Both achievements would have not been possible in a similar intuitive and transparent manner based on classical linear statistical models.

Teaching ML to doctoral students next to the classical education in inference statistics would not only provide PhD students with the tools to model complex, non-linear relationships, but also would sensitize them to the importance of out-of-sample testing and cross-validation. Understanding the philosophical difference between these two modeling cultures (see Breiman, 2001) is of great value to any empirical researcher and can be taught with a minimum of mathematical formalities.

A good starting point to teach ML is the R script provided in Hilbert et al. (2021). Many important steps of ML, such as data preprocessing, modeling, and evaluation, are performed step-by-step using a data set, making them easy to follow. This way ML can be explored in self-study by playing around with the different steps in the R script using the provided or other dataset, or a course can be built on this guide, especially when methods of iML that allow an easy access to ML, as chosen in this paper, are implemented.

REFERENCES

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, *17*(4), 267–321. https://doi.org/10.1007/s10887-012-9081-x

Hattie, J., & Yates, G. C. R. (2013). *Visible learning and the science of how we learn* (1st ed.). Routledge. https://doi.org/10.4324/9781315885025

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D., & Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*, *9*(3), Article e3310. https://doi.org/10.1002/rev3.3310

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Eds.). (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project*. Springer. https://doi.org/10.1007/978-1-4614-5149-5

Liebendörfer, M., Göller, R., Biehler, R., Hochmuth, R., Kortemeyer, J., Ostsieker, L., Rode, J., & Schaper, N. (2021). LimSt: Ein fragebogen zur erhebung von lernstrategien im mathematikhaltigen studium [LimSt: A questionnaire to collect learning strategies in mathematics-related studies]. *Journal Für Mathematik-Didaktik*, *42*(1), 25–59. https://doi.org/10.1007/s13138-020-00167-y

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable.* https://christophm.github.io/interpretable-ml-book/

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393