Towards Real Interpretability of Student Success Prediction Combining Methods of XAI and Social Science

Lea Cohausz
University of Mannheim
lea@informatik.uni-mannheim.de

ABSTRACT

Despite calls to increase the focus on explainability and interpretability in EDM and, in particular, student success prediction, so that it becomes useful for personalized intervention systems, only few efforts have been undertaken in that direction so far. In this paper, we argue that this is mainly due to the limitations of current Explainable Artificial Intelligence (XAI) approaches regarding interpretability. We further argue that the issue, thus, calls for a a combination of AI and social science methods utilizing the strengths of both. For this, we introduce a step-wise model of interpretability where the first step constitutes of knowing important features, the second step of understanding counterfactuals regarding a particular person's prediction, and the third step of uncovering causal relations relevant for a set of similar students. We show that LIME, a current XAI method, reaches the first but not subsequent steps. To reach step two, we propose an extension to LIME, Minimal Counterfactual-LIME, finding the smallest number of changes necessary to change a prediction. Reaching step three, however, is more involved and additionally requires theoretical and causal reasoning - to this end, we construct an easily applicable framework. Using artificial data, we showcase that our methods can recover connections among features; additionally, we demonstrate its applicability on real-life data. Limitations of our methods are discussed and collaborations with social scientists encouraged.

Keywords

Educational Data Mining, Student Drop-Out Prediction, XAI, Explainability

1. INTRODUCTION

Educational Data Mining (EDM) and in particular its subfield of student success and dropout prediction has gained prominence in recent years due to the increased digital education data availability and because the prediction of students' successes and struggles poses an important real-life problem. Accordingly, a multitude of studies exist testing various Machine and Deep Learning techniques on different data with some achieving remarkable accuracy and F1-valuesof more than 80 or even 90% regarding drop-out or success prediction [19, 13, 14, 6, 11]. This seems impressive and Xing & Du [17] write that individual drop-out probabilities can be used to "provide stronger and prioritized intervention to these students as a way of personalization" (p. 558). However, simply using the predictions will not enable one to do that. These only allow us to know *who* is likely to drop out; but in order to do anything with this prediction, we ought to know *why* the prediction has been made. Understanding why a prediction is made, is the topic of another timely topic in computer science, Explainable Artificial Intelligence (XAI). There, the *why* is split up further in *global* and *local* feature importance. With global, we mean what features are generally considered important by the model for predictions. This is valuable information so that we can control that features which would lead to a biased system discriminating against certain populations or features mistakenly included do not have an impact. A local explainability of a prediction, in contrast, relates to the importance of features regarding a specific person's prediction. This is important when we aim to use our predictions to help and advice a student predicted to be at risk as it allows us to understand what features contribute to their prediction specifically. Providing a basis to construct a personalized intervention system is the aim of this paper - thus, we work with local explainability.

The importance of XAI in EDM seems obvious and we are not the first to think so. Chitti et al. [5] heavily advocate a use of XAI techniques in future studies lamenting a lack thereof in current research. Alhamri & Alharbi [2] investigated the use of XAI and explainability techniques in performance prediction and came to the conclusion that few studies focus on it, and that those doing so merely focus on global explainability and do not employ techniques offered by XAI, instead simply using standard techniques as Decision Trees and looking at the generated rules. Indeed, it seems that with one notable exception [3], hardly any current research includes local explainability.

At first, it seems surprising that XAI techniques have not been employed more frequently as comparatively easy to use and model agnostic methods to extract important features exist[1]. However, we will see in this paper, that employing

L. Cohausz. Towards real interpretability of student success prediction combining methods of XAI and social science. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 361–367, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6853069

[1]SHAP for global, and LIME for local explainability [10,

XAI techniques off-the-shelf still leads to rather empty explanations. While knowing the important features explains why a prediction has been made, we also need to be able to interpret why and how these features matter at all. To better distinguish between those concepts, we argue for a distinction between *explainability* - which in this paper will relate to explaining the decision of the model, i.e. knowing the important features - and *interpretability* - which in this paper will relate to being able to interpret why and how features are important. Current XAI methods, as we will demonstrate later on, do not allow for interpretability.

Upon showing this, the objective of this paper is to provide a way of reaching interpretability as a basis for constructing personalized interventions. To this end, we will turn to another discipline, social science, that is traditionally well-equipped to deal with causal mechanisms and combine the predictive abilities of Machine Learning with the theory and causal analysis tools of social science. More precisely, we will:

- after a brief introduction to LIME, introduce a stepwise model of interpretability that will be illustrated with an example showing that XAI alone only achieves explainability but not interpretability.
- argue for a combination of XAI with social science, allowing us to gain full interpretability to construct a personalized intervention system. To this end, we will introduce a pipeline employing techniques from both disciplines.
- evaluate the approach on artificial data, so that we can test whether causal relationships are recovered; and show the applicability of the approach on real-life data.
- discuss the limitations of the approach and call for more collaborations among social scientists and computer scientists.

2. LIME

The aim of a personalized intervention system is both to provide a student with an idea of why they are predicted to struggle and what can be done to change this; and to provide pointers to those concerned with tailoring programs to help students. The requirements are, thus, to know what exact features contribute to a specific student's prediction (local explainability), which features' values would need to change in what way to change the prediction, and what underlying causal mechanisms are at work.[2] Note that when we consider causal mechanisms, we stray away from what XAI and purely data-driven approaches can provide. XAI techniques are concerned with explaining a model's decision but when we are interested in the causal mechanisms behind the important features we a) can no longer use XAI off-the-shelf and b) simultaneously make the assumption that the features important for the model also carry importance in real-life. The latter assumption should be kept in mind as it is not necessarily a given. Nonetheless, XAI can serve as a valuable basis for interpretability. Due to our focus on local interpretability, we choose to employ LIME. LIME is an acronym for Local Interpretable Model Agnostic Explanations which - as the name says - works for every model and finds local explanations for each instance. The basic idea is to randomly sample n feature vectors around the instance we want to explain given the normal distribution and to then weigh these new instances according to the distance to the instance we want to explain. Furthermore, the predicted label for each sample is obtained by feeding the feature vector into the model (note, that this works regardless of the specific method making it model agnostic). Based on this, Lasso regression is employed on the generated data with the predicted labels being the dependent (or target) variable. This allows us to extract the k most important features for the prediction [15].

3. A MODEL OF INTERPRETABILITY

One of the reasons why XAI techniques have not been used much may relate to their limitations regarding actual interpretability [9, 8, 1]. To illustrate our argument, consider Figure 1. It shows a Directed Acyclical Graph (DAG) of factors influencing whether a course, C1, is completed. We can observe whether courses C2, C3, and C4 are taken in parallel or not. These are our observable variables that could be used as features in a ML model. Taking these classes in parallel does not influence the completion of C1 directly. However, they do so indirectly through latent factors we cannot observe. Courses C2 and C3 compliment the contents of C1 well and taking them in parallel increases the competences required to complete C1 which then increases the probability of finishing the course. C4 is not related to C1 regarding content and thus does not contribute to competence important to complete C1. All three classes C2, C3, and C4 contribute to the workload, though. Having a high workload decreases the probability of finishing C1. Imagine now that for a student, Alice, the drop-out probability for C1 is predicted to be high. In order to fully leverage on this prediction, we should walk through each of the three steps of interpretability.

1. *Understand which features matter.* This means that we know which features matter for a person's prediction (local explainability). In our example, this means that the most important features regarding Alice's prediction are revealed to be the parallel taking of C2, C3, and C4. Furthermore, we know at this step that all three have a negative impact on completing C1. This is good to know but knowing the direction and impact of features is not equal to knowing what has to change in order to change the prediction. Should Alice not take any of these classes in parallel?
2. *Understand what would need to change to change the prediction.* In other words, we are looking for coun-

15].

[2]Note that while with global XAI, we often strive to exclude features that could be discriminating (e.g. age, gender, ...), local interpretability and the task of providing personalized intervention systems allows us to use those features as (a) people are not selected into programs because of it and more importantly (b) because we can try to mitigate difficulties certain populations have. If we identify, e.g., that some older students struggle, we could think about the mechanisms behind this, (e.g. older students generally have other responsibilities such as jobs and family competing for time), and and try to find solutions (e.g. provide flexible timetables, childcare on campus, ...).

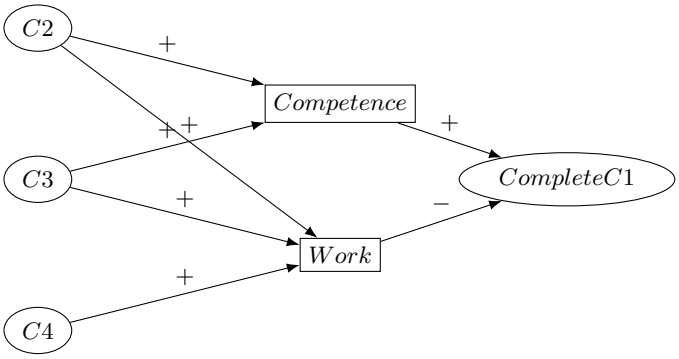

Figure 1: An example of causal modelling explaining factors influencing whether Course 1 (C1) is dropped or not.

terfactual explanations. In our case, we want to know whether not taking one or a combination of the classes will lead to a different prediction of our model. This can provide sensible advice.

3. *Understand the causal relationships among features and latent factors.* This refers to a causal understanding of features and latent factors. In this last step, we try to uncover the DAG (as shown in Figure 1) by theorizing about latent factors and testing whether the observations support this. We aim to understand that the courses influence latent factors competence and workload and which course influences which factor in which way. Not only does this lead to the best intervention for Alice; moreover, we can use this knowledge to construct better programs for all students.

If we use LIME off-the-shelf, we can extract the k most important features per person but counterfactual explanations are not provided. Thereby, we reach the first but not subsequent steps of interpretability. This limitation of XAI particularly regarding counterfactual explanations has been addressed and sometimes dealt with by other scholars as well [9, 8, 1]. Given these limitations, it is, therefore, maybe not surprising that few scholars saw it worth to employ XAI. How can we reach the other steps of interpretability and find a basis for personalized intervention? We argue that in particular reaching the third step calls for turning towards and employing techniques of social science. Firstly, we need an understanding of the concept of counterfactuals and aim to extend LIME in that way. Secondly, we need a theory-, instead of data-driven approach to explore causal mechanisms. Social science is well equipped for this task.

3.1 Reaching Step 2: MC-LIME

Reaching step 2 is easily possible as it rather naturally extends the idea of LIME, but requires an understanding of the concept of counterfactuals common in social science. In short, we attempt to answer the question of what would have happened regarding the outcome (the prediction) if the treatment (the features' values) had been different. In our case, we consider the k features extracted by LIME that have a *positive* impact on the drop-out probability (i.e. make it more likely that someone drops out). Then, we check for the smallest subset of these features that - when changing their values - changes the prediction and return the features and changed values[3]. We achieve this by iteratively changing one feature's value; if this never leads to a prediction change, we check for all combinations of two features etc. Because we look for the smallest necessary change, we call this approach Minimally Counterfactual LIME (MC-LIME). If multiple subsets of the same size exist, but we only want a certain number, we can select to receive the c changes that lead to the largest difference in the output probability. This procedure is straightforward for binary variables where we can simply use the complementary value. For categorical and ordinal features, we propose iterating through all possible values; all values, for which a change is reported, are stored - the highest change counts towards the selection of the top c features. For count features, we propose shifting the value a standard deviation towards the mean, so that the change is large enough to make a substantial difference. The resulting subset tells us what would minimally need to change in order to change the prediction and how this change would need to look like.

3.2 Reaching Step 3: A Causal Analysis

While this information is already very important, it is not enough to provide good interventions and to potentially construct programs, though. In order to know why features matter - for the model but hopefully also overall - we propose to use *all* features LIME returns (positive and negative impact) as a basis for a deeper analysis. For this theory-driven approach, we propose to follow the steps:

1. Extract all features and their impacts and use it to cluster people into groups. Therefore, we only work with a subset of all extracted variables that are known to be relevant for a set of students thereby simplifying the model while at the same time assuring that we use relevant features.

2. For the demographic (and if available social and psychological) features, e.g., age, having a student job, living closer or further away from university, look for social science studies that investigate their effect on drop-out and let it inform you on causal mechanisms. If you find that other features could also be important, add them.

3. For features specific to your domain, e.g. the courses offered, try to understand what they are about and how this could influence the outcome variable, i.e. the drop-out.

4. Begin drawing a DAG and consider the following questions: (a) Is a connection between two variables direct or does it go through a latent variable we cannot observe? Does the latent variable mediate the effect? (b) Is there an actual relationship between two variables or are they confounders meaning that a third variable effects both? (c) Does a third variable moderate the effect between two variables? (d) Is the effect linear or quadratic?

[3]Note that this is very similar to LIME's understanding of feature importance.

5. Model a regression formula according to your DAG and run the regression on the training data with success or drop-out being the dependent (or target) variable.

6. Check the effects of the terms of your formula. What is significant? Does the direction conform to your theoretical considerations?

7. Construct personalized interventions according to social science theory and in combination with the insights of step 2.

4. EVALUATION

In order to demonstrate the pipeline and its applicability, we test our approach on an artificial and a real-life set of data.

4.1 Data

Artificial Data. As, typically, we do not know the real causal mechanisms and can only make informed guesses considering the existing literature and our own reasoning, we test MC-LIME and our causal framework on artificial data. Regarding MC-LIME we can test whether it returns the feature subset that is intended to make a difference. It is, of course, not very telling to use this data on our causal framework as we know the causal mechanisms we decided on. However, it is still valuable to see whether we can recover the intended effects and their directions as specified in data generation. Our data consists of the target variable drop-out and 26 other binary features. Of these 26 features, when they are set to 1, eight have no effect, three have no direct effect but do have one when combined with other variables, two have a negative impact on drop-out that reverses when combined with other variables, three have a positive impact that reverses when combined with other variables, five have a negative effect and five have a positive effect. The first row of Table 1 summarizes this; a plus indicates a positive effect on drop-out, a minus a negative effect. The number of symbols represents the strength of the effect, as each causal relationship was given a weight by which the probability of a drop-out changes. We created 10,000 instances by randomly sampling features. The drop-out value was determined by the sum of of the weights of the non-zero variables. If the sum was 0.5 or higher, we assigned a 1, else we assigned a 0. This resulted in 30% of instances having assigned a 1.

Real-Life Data. In order to demonstrate the process on real-life data, we gathered information on a mandatory first-year theoretical computer science course - that we will call C1 - part of a three-year Bachelor degree at the University of Mannheim, Germany.[4] We have information on all students that registered for this course between 2010-2020 and try to predict who will drop out. Note that students who failed or dropped out once and then registered in subsequent years may appear in the data twice. Our data contains 1,738 instances. Furthermore, even though the course is meant to be taken in the first year, many students take it later. The data contains seven demographic features, and 160 features on high school results, previous courses taken, previous results and drop-out behavior, and classes taken in parallel. To understand our data structure, consider a course A. This course has four features assigned to it: whether is is taken in parallel to C1, whether the student failed it, whether the student dropped out, or whether the student passed. Note that a student can have 1 assigned to several of these features, if, e.g., a person first dropped and then passed course A. Again, remember that this only encompasses the information we have at the time when the student registers for the course we are predicting on. In total, we consider 30 courses. Table 2 provides summary statistics of the data.

[4]The data was k-anonymized prior to analysis to ensure privacy.

4.2 Step 1: LIME

For both sets of data, we predicted the drop-out using several methods: Support Vector Machine (SVM), a simple Deep Neural Network (DNN), Naive Bayes, Decision Tree, and Random Forest. Then, we selected the model leading to the highest F1-value and accuracy in the test data. For the artificial data, this was the DNN with an accuracy of 99.5% and a F1-value of 0.99. For the real data, it was the SVM with an accuracy of 87.12% and an F1-value of 0.9. Having selected the best model, we extracted the ten most important features and their directions for each instance of the test data. We only considered those instances for which drop-out was predicted, as these are the ones we are most interested in.

Artificial Data. 318 of the 1,000 test instances were predicted to drop out. Table 1 shows which features where extracted at least for one instance. We can see that feature $V8$ was not considered important at all, even though it is supposed to have a negative effect on drop-out. In contrast, $V20$ was extracted once, even though it should not have an effect. Furthermore, $V17$ and $V18$ were not extracted; these features do not have an effect on their own, but do when combined. The most extracted features were $V2$, $V11$, $V12$, and $V13$ which were extracted for each instance, followed by $V5$ (316), $V1$ (312), $V6$ (284), and $V3$ (213). Note that this does not mean that for all those predicted to drop out, each of these variables was set to 1 or had a positive impact as the table also includes features that have a negative impact on drop-out. As a matter of fact, the extracted directions of the effects are correct for all extracted features that upon being set to 1 are supposed to have a positive or negative effect on drop-out. For those features for which the direction of the effect changes upon combination with others, we can see that the reverse effect is extracted. This shows the limitations of LIME and, therefore, the importance of our remaining steps.

Real Data. For 26 instances the label drop-out was predicted. In total, 25 important features were extracted; of these, six only appeared once. The most frequently extracted features were whether a person planned to take the exam on the first date or in the resit (26)[5], the study year (26), the age (26), whether two other first year classes were taken in parallel (24 each), whether one of these first-year courses had been dropped before (20), and whether a second year course had been passed (12).

[5]Students have the opportunity to decide between taking the exam right after the lecture period or two months later; the latter is known as the resit date.

Table 1: Artificial Data - variables, their effects, and what could be recovered using LIME and regression on extracted features.

Effects	Pos. Effect	Neg. Effect	Effect in Combination	No Effect
Variables	V1(++), V2(++), V3(++), V4(++), V5(++)	V6(−−), V7(−), V8(−), V9(−), V10(−)	V11(−−)+V12(−−)+V13(No): +++; V14(+)+V15(+)+V16(+): − − −; V17(None) + V18(None): ++	V19-V26(No)
Recovered (LIME)	V1(+), V2(+), V3(+), V4(+), V5(+)	V6(−), V7(−), V9(−), V10(−)	V11(+), V12(+), V13(+), V14(−),V15(−), V16(−)	V20(−)
Changed Prediction (MC-LIME)	V1, V2, V3, V4	V6, V9	V11, V12, V13	V20
Recovered (Regression)	V1(+), V2(+), V3(+), V4(+), V5(+)	V6 (−), V7 (−), V9(−), V10(−)	V11(−)+V12(−)+V13(None): +; V14(+) + V15(+) + V16(None): − ; V17(None) + V18(None): +	V20(−)

Table 2: Summary statistics of real-life data regarding the first year course C1.

Variable	Key Statistics
Age	20.49 (min: 16, max: 36)
Year	1.7 (min: 1, max: 5)
N of Attempts	1.3 (min: 1; max: 3)
Gender	female: 18.35%, male: 81.65%
Nationality	domestic: 83.77%
Domestic HS Degree	91.48%
Drop-Out of C1	41.49%

4.3 Step 2: MC-Lime

Artificial Data. We now selected those important features that have a positive effect on drop-out for each instance and iteratively changed the values. For 302 instances, it was enough to change a single value to change the prediction; for 14 of these, there was only one feature which managed to change the prediction on its own. Table 1 displays what features changed the prediction on their own for at least one instance. 14 instances needed two changes, the remaining three changes. The feature most often leading to a change when assigned a different value was $V11$ (291), followed by $V13$ (288), 12 (287), and $V2$ (112). Interestingly, $V20$ also changed the prediction on its own once. Several variables could not change the prediction on their own. Apart from $V5$, though, these are only variables that have a small impact on the drop-out rate in comparison.

Real Data. Proceeding in the same fashion, we found that 14 instances only needed a change in one feature to change. This rose to 19 when we considered changes in features referring to the same course as just one. There were 10 instances for which only one specific feature changed the prediction. Two instances needed two changes, the remaining instances three or more. 16 features changed the prediction on their own for at least one instance. The instances most often leading to changes were relating to two of the three classes extracted as important before (16, 13), the variable indicating the date (16), the age (8), and semester (6). Of course, a person cannot change their age upon learning that this contributes to the prediction. However, universities can identify causal mechanisms explaining the importance of age and then construct specialized offers. In order to be able to this, we of course need to continue with step 3.

4.4 Step 3: Regression and Modelling

Artificial Data. For the artificial data, we simply used all the extracted features and our knowledge [6] about the data generation to construct the logistic regression formula. Then, we entered this together with the training data in a logistic regression. The last row of Table 1 shows the recovered effects. All variables entered into the regression were significant (then they were given the sign of the direction of their effect in the table) apart from $V16$ and $V11$ - even $V20$ (albeit only on the 5%-level) which means that by chance in data generation, more instances got the label drop-out assigned which also received a 1 in this variable. $V16$ was positive but not significant, even though it should be. $V11$ was correctly identified as no longer being significant once combined with the other two variables. All effects now also had the correct directions. Interaction terms of $V11$, $V12$, and $V13$ and $V14$, $V15$, and $V16$ were also significant and had the correct direction. We can see that reasoning about and investigating causal mechanisms made it possible to recover most effects and their directions.

Real-Life Data. For the real-life data, as explained above, we first clustered our test instances using k-Nearest Neighbor based on the features extracted so that we only focus on features relevant for this set of students. We chose $k = 3$ upon visual inspection. For illustrative and space reasons, we will focus on the largest of the clusters containing twelve instances. We used all features that were extracted for more than one instance in the cluster (Table 3). The only two features that are not specific to our setting are the variables age and nationality. Therefore, we consult the literature on these variables. For age, scholars are divided. While some studies stress that older students are generally more successful and achieve higher grades, others find that advances in age can also be seen as a positive predictor for drop-out [16, 18, 4]. The former is generally attributed to older students being more certain of their goals and having an increased focus; the latter is often attributed to having other important parts of life such as a family or a job. Based on this, we theorize that those being a little older are influenced by the former, and those who are much older by the latter; thus, for the regression, we include age and age^2. Of course, that age has been selected in the first place, could also be due to the fact that those students taking the class later in their studies are older and may regularly struggle with courses. To account for this, and because it is also extracted as a feature, the number of years one has studied is also included. For domestic nationality, the likely reason

[6] Therefore also entering $V17$ and $V18$ again.

for the negative effect LIME implies is that the degree is in German which creates a language barrier to non-German speaking students [12, 7]. This might be mitigated when a student already received their high school diploma in Germany; thus, we enter this variable in an interaction term, even though, it is not extracted as an important feature. The other features are specific to our setting. We argue that having failed the course before leads to a decreased probability of dropping out because students have already completed the course before. Writing the exam on the resit date leads to an increased probability of dropping out because the exam is written almost two months after the end of lectures meaning that a) students may have not paid enough attention on this course during the lecture period and b) students may have forgotten important information in the meantime. We argue that having passed courses C2, C3, and C4 leads to a decreased drop-out probability, because these courses have connected contents and require a similar skill-set. Those who did not struggle much with these courses can then also complete this one. Likewise, having struggled in these courses leads to a higher drop-out probability. Furthermore, we argue that taking C2 and C3 together with our queried course C1 - as intended by the study program - may lead to a very high workload; thus, we also include an interaction term of these. Table 3 shows our results, the middle column summarizes the results of the logistic regression without interaction or quadratic terms, the right-hand column the results including these terms. The symbol "+" indicates a positive effect on the drop-out probability, symbol "-" a negative one. One "+" indicates an effect on the 10%, two on the 5%, and three on the 1% level; a "No" indicates no significant effect. We can see that the effect for age - which is at first positive - reverses when age^2 is added with higher age now leading to a decreased drop-out probability on average, though age^2 is not itself significant most likely due to the small sample size. Similarly, the non-significant effect for domestic students and high school diploma may be due to that. The study year does not matter, but the date greatly matters with taking the exam at a later date leading to a higher probability on average. Having failed the course before, leads to a smaller drop-out probability, but having dropped it to a larger one. Only taking C3 in parallel seems to have an effect on its own. C2 has no effect. For C3 we see that those who dropped this also have a higher probability of dropping C1; having passed C3 also means that C1 will likely be completed. For C4, even not having passed already leads to a smaller probability of drop-out. Having passed all three courses also leads to a smaller drop-out probability. What do we take away from this? Generally, we should investigate whether the age effect that comparatively old and young students struggle persists across the overall program. If so, we should think about how to help these age groups. Furthermore, considering this course in particular, we should encourage the students to not choose the resit date. We should also identify students who have struggled with the courses of similar content before and offer increased assistance and attention to them. How to best do this, is something social science may also help us with. Finally, we combine our insights of step 2 and 3 for each student to tailor the best intervention.

5. CONCLUSIONS, LIMITATIONS, & FUTURE WORK

Table 3: Regression results of real-life data.

Variable	Without Interaction	With Interaction
Age, Age^2	+	−, No
$Year$, $Year * Age$	No	No, No
$Date$	+++	+++
$Domestic$, $Domestic * HS$	No	No, No
$failC1$, $dropC1$	No, +	−, +
$parC2$, $parC3$, $parC4$, $parC2 * parC3$	No (all)	No, −, No, No
$failC2$, $dropC2$, $pasC2$	No(all)	No(all)
$failC3$, $dropC3$, $passC3$	No, +, No	No, +, −
$failC4$, $dropC4$, $passC4$	−, No, −−	−, No, −−
$passC2 * passC3 * passC4$		−

In this paper, we addressed the issue of a lack of XAI in EDM. More specifically, we attempted to provide a framework enabling us to use the information extractable from predictions as a basis for a personalized intervention system. To highlight the challenges and requirements, we came up with a step-wise model of interpretability where step 1 means identifying important features, step 2 identifying the minimal set of value changes to change the prediction, and step 3 identifying causal mechanisms. We described methods to reach each of the steps and evaluated them on an artificial and real-life dataset showing the applicability. Our results on artificial data showed that the method works well when we correctly theorize about causal mechanisms. Of course, we may not always able to do that. This is a limitation of our work which, in general, provides no "one size fits all"-formula, but needs to be adjusted for different settings. Furthermore, our methods can certainly be further refined but we hope that our step-wise model of interpretability provides a good orientation. A third limitation is that we do not show an actual application of our method in a real-life setting meaning that we cannot evaluate whether the conclusions derived from our analysis benefit the students in practice. This is a future endeavour. When using predictive systems in practice, we would like to once again stress that this should be made clear to students and should be very transparent. Finally, we would like to argue for an increased collaboration among social and computer scientists. Whereas computer scientists are typically experienced with predictions and deriving knowledge from data, they lack experience when it comes to theory-driven approaches and causal analysis. Social scientists, in contrast, typically do not work on predictions but are knowledgeable regarding statistical tools to uncover causal mechanisms and deriving models from theory. While these differences in approaches are prone to hinder collaboration, for this task it will greatly benefit both disciplines. Furthermore, while social science informs our models, it can also gain new insights through large-scale predictions and deriving information from data.

6. REFERENCES

[1] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[2] R. Alamri and B. Alharbi. Explainable student performance prediction models: a systematic review. *IEEE Access*, 2021.

[3] M. Baranyi, M. Nagy, and R. Molontay. Interpretable deep learning for university dropout prediction. In *Proceedings of the 21st Annual Conference on Information Technology Education*, pages 13–19, 2020.

[4] R. Chen. Institutional characteristics and college student dropout risks: A multilevel event history analysis. *Research in Higher education*, 53(5):487–505, 2012.

[5] M. Chitti, P. Chitti, and M. Jayabalan. Need for interpretable student performance prediction. In *2020 13th International Conference on Developments in eSystems Engineering (DeSE)*, pages 269–272. IEEE, 2020.

[6] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro. Student dropout prediction. In *International Conference on Artificial Intelligence in Education*, pages 129–140. Springer, 2020.

[7] S. Evans and B. Morrison. Meeting the challenges of english-medium higher education: The first-year experience in hong kong. *English for Specific Purposes*, 30(3):198–208, 2011.

[8] M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. *arXiv preprint arXiv:2103.01035*, 2021.

[9] M. T. Keane and B. Smyth. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In *International Conference on Case-Based Reasoning*, pages 163–178. Springer, 2020.

[10] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[11] R. Manrique, B. P. Nunes, O. Marino, M. A. Casanova, and T. Nurmikko-Fuller. An analysis of student representation, representative features and classification algorithms to predict degree dropout. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 401–410, 2019.

[12] L. Morrice. Refugees in higher education: Boundaries of belonging and recognition, stigma and exclusion. *International Journal of Lifelong Education*, 32(5):652–668, 2013.

[13] B. Prenkaj, P. Velardi, G. Stilo, D. Distante, and S. Faralli. A survey of machine learning approaches for student dropout prediction in online courses. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[14] L. Qiu, Y. Liu, Q. Hu, and Y. Liu. Student dropout prediction in massive open online courses by convolutional neural networks. *Soft Computing*, 23(20):10287–10301, 2019.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[16] T. M. Spitzer. Predictors of college success: A comparison of traditional and nontraditional age students. *Journal of Student Affairs Research and Practice*, 38(1):99–115, 2000.

[17] W. Xing and D. Du. Dropout prediction in moocs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*, 57(3):547–570, 2019.

[18] D. Yasmin. Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. *Distance Education*, 34(2):218–231, 2013.

[19] H. Zeineddine, U. Braendle, and A. Farah. Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, 89:106903, 2021.