When Probabilities Are Not Enough — A Framework for Causal Explanations of Student Success Models

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

Student success and drop-out predictions have gained increased attention in recent years, connected to the hope that by identifying struggling students, it is possible to intervene and provide early help and design programs based on patterns discovered by the models. Though by now many models exist achieving remarkable accuracy-values, models outputting simple probabilities are not enough to achieve these ambitious goals. In this paper, we argue that they can be a first exploratory step of a pipeline aiming to be capable of reaching the mentioned goals. By using Explainable Artificial Intelligence (XAI) methods, such as SHAP and LIME, we can understand what features matter for the model and make the assumption that features important for successful models are also important in real life. By then additionally connecting this with an analysis of counterfactuals and a theory-driven causal analysis, we can begin to reasonably understand not just *if* a student will struggle but *why* and provide fitting help. We evaluate the pipeline on an artificial dataset to show that it can, indeed, recover complex causal mechanisms and on a real-life dataset showing the method's applicability. We further argue that collaborations with social scientists are mutually beneficial in this area but also discuss the potential negative effects of personal intervention systems and call for careful designs.

Keywords: student drop-out prediction, xai, explainability, interpretability

1. Introduction

Educational Data Mining (EDM) and, in particular, its subfield of student success and drop-out 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 (black-box) Machine and Deep Learning (ML, DL) techniques on data from diverse sources, with some achieving remarkable accuracy and F1-values of more than 80 or even 90% regarding drop-out or success prediction (Zeineddine et al., 2021; Prenkaj et al., 2020; Qiu et al., 2019; Del Bonifro et al., 2020; Manrique et al., 2019). Generally, predictions can either be made on the course- or degree-level; i.e., we can either predict whether a student is likely to drop-out/succeed in a given course or we can predict whether a student is likely to drop-out/succeed in their study program (Prenkaj et al., 2020; Del Bonifro et al., 2020). For the former, either classical tabular data on demographics, previous courses, current grades, etc. is used to predict with ML or DL models (Prenkaj et al.,

2020); or clickstream data is used (typically in settings of Massive Open Online Courses) either directly using Convolutional Neural Networks (CNN) (Qiu et al., 2019) or after preprocessing then again in an ML or other DL models (Xing and Du, 2019). For degree-level data, typically tabular data is used in diverse ML and DL models (Del Bonifro et al., 2020; Manrique et al., 2019). Due to the class imbalance, it is considered good practice to report F1-value (Prenkaj et al., 2020) and, as already mentioned, even then, numbers of more than 75% are not at all unusual (Manrique et al., 2019).

This seems impressive and Xing and Du (2019) 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). Indeed, predictions in EDM are typically considered more of a means towards intervening in a sensible way or constructing study programs or courses instead of as an end of its own. The hope typically is to reduce drop-out and failure, which benefits many stakeholders, from students to universities to the state. The question, though, is how we get from predictions to interventions. Rather obviously, predictive models only tell us *who* but not *why* someone is likely to drop out. To make this problem even more apparent, consider the following example: Imagine that we have two students, Alice and Bob, both thinking about taking a course and a model that predicts that Alice will probably finish the course and that Bob will not. What does this mean? Should Bob not take the course? Why is his prediction bad? Can he change something so that his prediction becomes positive? Additionally, we may not only be interested in personalized interventions for the students but also in detailed feedback for the instructors (e.g., telling them what topics are particularly challenging, when people are known to drop-out) and support for administrators and those designing study programs (e.g., telling them that people with certain characteristics tend to struggle because of specific reasons, telling them about the effects of schedules). Ideally, the different stakeholders can then respond to this (e.g., by spending more time on topics or providing help to students with specific characteristics). This makes it clear that we need more than just predictive models.

A dominant notion right now is that we could use post-hoc Explainable Artificial Intelligence (XAI) to gain information and construct interventions from predictions stemming from powerful black-box models (Chitti et al., 2020; Alamri and Alharbi, 2021; Conati et al., 2021). XAI is typically used to explain a model's predictions. Explanations can either be on the local or global level (Adadi and Berrada, 2018). With global explanations, we are interested in what features are generally considered important by the model for predictions. This is valuable information so that we can control features that would lead to a biased system discriminating against specific populations or make sure that features mistakenly included do not have an impact. In contrast, a local explanation of a prediction relates to the importance of features regarding a specific person's prediction. This is most important when we aim to use our predictions to help and advise a student predicted to be at risk, as it allows us to understand what features contribute to their prediction specifically. Thus, we are most interested in local explanations. Interventions based on these explanations can then be constructed in one of two ways: Either students can be clustered according to their explanation and then a manually designed intervention, e.g., based on what the instructor believes from experience to be a sensible intervention in this case, is assigned to these students (Conati et al., 2021). Or the intervention is directly derived from the explanation (Mu et al., 2020). In particular, doing the latter seems desirable as the hope is that we can learn something from the predictive model about relationships in the data. However,

using XAI in EDM, in general, is (save for a few examples) not yet very common, and even less so is using XAI to construct interventions in EDM (Chitti et al., 2020; Alamri and Alharbi, 2021). In part, this may have to do with both EDM and XAI being rapidly developing and rather new research areas where methods are still being developed. However, we believe that simply employing XAI on predictive models does not provide the kind of information we can use for interventions. 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 of the model* — which in this paper will relate to explaining the decision of the model, i.e., knowing the important features for the prediction — and *interpretability of the world through the model*[1] — which in this paper will relate to being able to interpret why and how features are important with the help of a model. Current XAI methods, as we will demonstrate in this paper, do not and cannot allow for interpretability but provide the basis. Therefore, the lack of existing research may also be due to these limitations. This, in turn, makes it necessary to create a framework that allows us to construct interventions from black-box predictive models despite the limitations. In this paper, thus, we will:

- Provide an introduction to XAI as as well LIME and SHAP, two prominent post-hoc XAI-methods. We will also discuss the few papers that have, indeed, used local explanations in EDM and briefly look at papers using XAI to construct interventions. Doing so will reveal why these methods are limited in their usability.
- State and discuss the requirements to construct interventions from predictive models and provide a discussion of causality.
- Provide a framework of how to meet the requirements. The framework will consist of a pipeline that uses insights from computer and social sciences.
- Evaluate the approach on artificial data so we can test whether causal relationships are recovered and show the approach's applicability on real-life data.
- Discuss how to construct interventions considering existing literature in EDM on interventions.
- Argue how this area of research could benefit from collaborations among social and computer scientists.

Therefore, this paper presents an extended version of a previously published conference paper Cohausz (2022). In difference to the conference paper, this version contains a deeper discussion of the XAI-methods and previous work in EDM using them as well as a section on interventions and intervention design in EDM. More details are given regarding the implementation and evaluation, and potential collaborations between social and computer scientists are discussed.

[1]Note that the terms explainability and interpretability are frequently and sometimes interchangeably used and hardly ever defined in the literature. Meske et al. (2022) defined interpretability as a property models can have, which means that no proxy-explanation of the model in the form of post-hoc XAI-methods is needed to understand the machine's reasoning. We look at it from another angle, saying that post-hoc XAI-methods do not offer interpretability - which is in accordance to Meske et al. (2022) — but that interpretability is nonetheless achievable — even for black-box models — as long as we can learn about the real-world connections.

2. XAI

XAI-methods are typically used to understand how a model reaches a decision, i.e., to explain the decision. For some models, it is not necessary to use any special algorithms to do so, as they can provide explanations by default. Decision Trees, e.g., generate rules that humans can look at and understand. Similarly, for (logistic) regression, we can look at standardized effect sizes, which show which features contribute to the decision in particular. In recent years, however, specialized, so-called post-hoc algorithms were invented aiming to explain the predictions of all models or a type of model (Lundberg and Lee, 2017; Ribeiro et al., 2016) and the topic of XAI as a whole gained popularity (Meske et al., 2022; Adadi and Berrada, 2018). This, of course, mostly has to do with the increasingly complex black-box models of Deep Learning. In Deep Learning and some standard Shallow Learning models, it is impossible to directly observe the effects of features[2] or the generated rules, and even if we did observe them, we might not be able to understand them. The objective of post-hoc XAI-methods, hence, is to try to generate humanly understandable explanations. These explanations typically aim to ensure no group is discriminated against, increase trust in the application, or check that no insensible features are used. Adadi and Berrada (2018) summarize the goals of post-hoc XAI as being able to justify, control, improve, and discover. However, the exact goal and scope of XAI remain debatable. We will now explain LIME and SHAP, two prominent XAI-methods, before returning to the question of what we want from XAI in general and what we want from it in EDM.
Generally, what both LIME and SHAP do is to return features (along with directions indicating how the features are important) that are important for the model's decision. Both methods essentially define a feature as important if a model returns a different prediction when the value of this feature for a given data instance is changed (Ribeiro et al., 2016; Lundberg and Lee, 2017). Thus, both SHAP and LIME belong to the perturbation-based XAI-methods.

2.1. LIME

LIME — an acronym for Local Interpretable Model-agnostic explanations — was introduced by Ribeiro et al. (2016). The idea behind LIME is to explain an instance's, $\mathbf{x}$, prediction by sampling n new instances $\mathbf{x}'$ close to the instance's feature vector and weighing them according to the distance to $\mathbf{x}$. For each new instance, we compute the model's prediction y'. The weighted instances and their predictions are used to train a simple white-box local model, usually either Ridge or Lasso regression, which mimics the original model. Features that influence the prediction more receive larger coefficients, and a user-defined number k of the most important features and corresponding coefficients is returned. More formally, the general idea behind LIME is to receive explanation $\xi(\mathbf{x})$ by minimizing:

$$\xi(\mathbf{x}) = argmin_{g \in G} \mathcal{L}(f, g, \pi_{\mathbf{x}}) + \Omega(g)$$

where f is the model we want to explain and g is a locally interpretable model of all possible models G with $\Omega(g)$ being a measure of complexity of the interpretable models.

Figure 1 shows an explanation for the prediction of a Titanic-passenger's survival created

[2]Throughout this work, we will use "feature" and "variable" interchangeably. The former is typically used in computer science, and the latter in social science.

with the well-known Titanic-dataset with k being set to 10 [3]. The Titanic-dataset is a classic dataset to try ML-methods on with the aim of predicting whether a person (i.e., a data instance) survived. The dataset consists of such variables as the sex of the passenger, the number of family members (family size), the passenger class (e.g., first class), or whether that passenger had a title such as, e.g., "Sir". The table in figure 1 ranks the extracted features by importance for the prediction, i.e., the passenger class and sex are more important for the prediction than family size or title. We can furthermore see what value the passenger has for each feature and whether this value is perceived as increasing the chances of survival (orange color). For example, being in the first class positively affects survival (the feature is Pclass and the value is 1), and being female (the feature is Sex and the value is 0) does too. The figure next to the table shows the rules that LIME has discovered. For example, the title-variable having a value larger than two has a positive effect (and as the title of the instance is three, it has a positive effect).

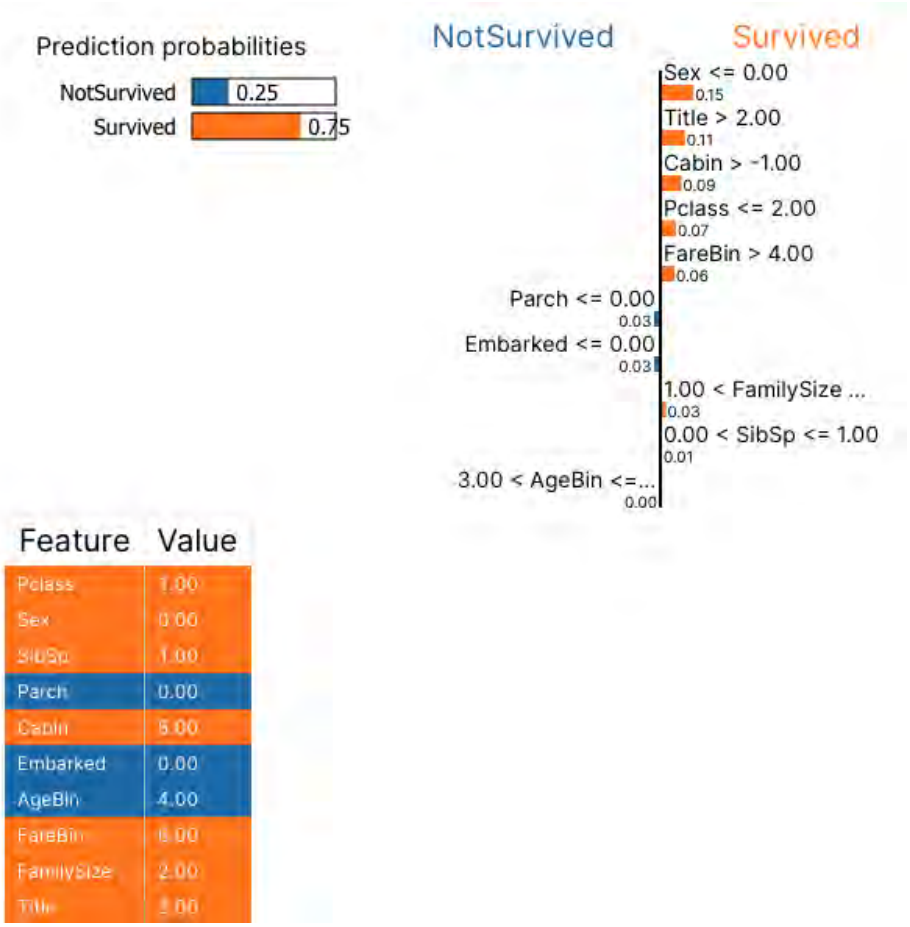


Figure 1: Example of a LIME explanation for a Titanic-passenger.

[3]We use the example from https://www.kaggle.com/code/yuu113/model-interpretation-shap-lime-with-titanic-data.

2.2. SHAP

SHAP (SHapley Additive exPlanations) is based on Shapley values known from game theory and LIME (Lundberg and Lee, 2017). Shapley values are typically used in cooperative game theoretic settings to allocate credit to players according to their contribution to the mutual profit. Likewise, SHAP allocates credit to features according to their relevance for the prediction. This is achieved by looking at the predictions using only a coalition of a subset of features and measuring the average marginal contribution for each feature across all coalitions. The way the Shapley-values are calculated varies for different SHAP-variants, and there exist SHAP-variants specific to the predictive model (e.g. TreeExplainer, GradientExplainer). We will now look at (and use) KernelSHAP as it is the model-agnostic variant.
If we want to explain instance $\mathbf{x}$, we sample a feature (or coalition) vector $\mathbf{z}'$ which is equal in length to $\mathbf{x}$ and consists of 1s in places where features are present and 0s in places where features are absent, i.e., we are masking some features. Features are weighed according to the kernel which is based on the coalition size:

$$\pi_{\mathbf{x}}(\mathbf{z}') = \frac{(M-1)}{\binom{M}{|\mathbf{z}'|}|\mathbf{z}'|(M-|\mathbf{z}'|)}$$

Here, M refers to the maximum coalition size (i.e., number of features) and $|\mathbf{z}'|$ to the number of features actually used in the coalition vector. Like for LIME, a local model is then trained and the explanation is defined as:

$$g'(\mathbf{z}') = \phi_0 + \sum_{i=1}^{M} \phi_i \mathbf{z}'$$

Again, g' refers to a local explanation model where ϕ corresponds to regression coefficients (Lundberg and Lee, 2017). So like for LIME, we use a linear model to predict on weighted samples. In difference to LIME, weights are computed using the kernel, which assigns the largest weight to large and small coalitions (many 1s or 0s, respectively) - which should provide us with the best idea of an individual feature's contribution. Remember that LIME assigns larger weights to vectors close to the input, so we can think about that in terms of assigning a large weight to large coalitions. Another difference to SHAP is that we do not specify how many features get extracted but only an upper limit. Additionally, and relevant for interpreting SHAP's output, SHAP displays the coefficients computed.

Figure 2 shows the explanation produced by SHAP for the same passenger that we also considered for LIME. Red features represent those having a positive impact on survival, blue otherwise. Again, we can see this passenger's values for each feature. Overall, the two methods produce rather similar results for this data instance. However, as already mentioned, SHAP displays the coefficients. -0.5914 corresponds to ϕ_0 and can be thought of as β_0 in a normal regression model. For each relevant feature, we can see the effect of the instance's feature value on survival, e.g., for the passenger's family size computed as $\phi_{FamilySize} * 2$. Again, this can be thought of as β_x in a regular regression.

In summary, we can see that both methods return information on individual feature importance and direction regarding an instance's prediction of the model. The question is, now, what

Figure 2: Example of a SHAP explanation for a Titanic-passenger.

we actually want from explanations and whether LIME and SHAP are capable of this.

2.3. What We Want From Explanations

Indeed, it is relatively unclear what constitutes a good explanation and what XAI should be capable of doing (Adadi and Berrada, 2018; Emmert-Streib et al., 2020; Meske et al., 2022; Langer et al., 2021). Miller (2019) as well as Keane and Smyth (2020) state that the explanations aimed to explain predictions or interventions resulting from the prediction to users should be of a contrastive type and need not be complete. Miller (2019) states that people generally ask why something happened instead of something else and then pick the parts of the explanation relevant to them; hence, explanations should be steered in that direction. While Meske et al. (2022) and Langer et al. (2021) state that the type of explanation (and their completeness) should depend on the stakeholders the explanation is aimed at, they also do not contest the evidence from social science that indicates that we tend to understand explanations in the form of counterfactuals. The idea behind using XAI to construct interventions is also based on the perception that we can learn something about what-if scenarios through explanations.
LIME and SHAP are not necessarily good explanations in this sense. They do not provide explanations that can directly be used as counterfactuals. For example, if the passenger had a different gender, it would decrease their chance of survival, but it would not necessarily mean that they would be predicted to die. Furthermore, both methods do not really enable people to pick explanations relevant to them. The next question is, though, what we want from explanations when we aim to construct interventions from them or in the field of EDM in general.

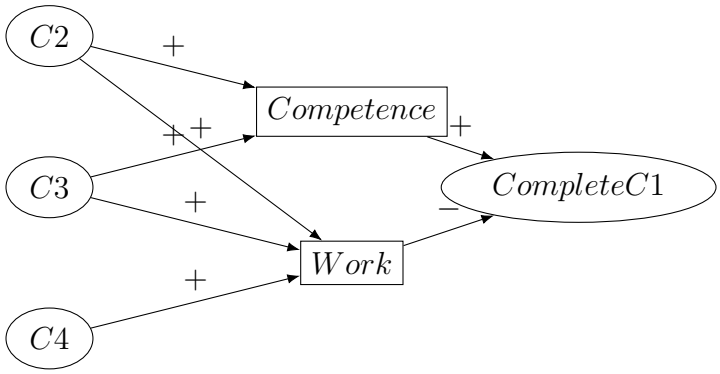
In fact, using XAI for interventions instead of justifying explanations is a wholly different scope. If we intend to use explanations for the construction of interventions, they need to be complete and correct. That XAI in EDM is often used outside the original scope of providing an explanation to justify a decision has also been discussed by Khosravi et al. (2022). They write that in EDM, XAI is supposed to do more than justify predictions; instead, it should provide feedback and a basis for changes. There, having incomplete (or wrong) explanations or a poor model leads to drastic errors (Khosravi et al., 2022). Hence, in this area, we need complete and correct explanations. The pressing question, then, is what constitutes as a complete and correct explanation and whether LIME and SHAP provide it. This is more difficult to answer. We might note that both methods return a set of important features but whether this set entails all important (or maybe more than the important) features is not immediately obvious. Further, neither method provides information on the specific relationship between variables as potentially used in the ML-model. For example, it could theoretically be the case that being a woman and being young both individually have a positive impact on survival but that being a young woman turns the effect around to a degree. This relationship would not be visible in LIME's and

SHAP's explanations as they only return information on individual variables. Note, furthermore, that such a relationship could lead to estimating wrong impacts of each individual variable. These aspects already point to LIME and SHAP not necessarily returning complete and correct explanations, but it still does not really say what exactly we would need. To get a better idea regarding this, we will now look at how XAI is currently used for interventions in EDM.

3. Interventions in EDM

Much predictive modeling research in EDM is ultimately aimed at producing interventions. In particular intelligent tutoring systems, writing or curriculum assistants try to provide personalized advice and interventions (Khosravi et al., 2022; Conati et al., 2021). While the usage of XAI in EDM is generally still limited (Baranyi et al., 2020), some works aiming to provide personalized interventions have used XAI (Conati et al., 2021; Khosravi et al., 2022; Mu et al., 2020; Hur et al., 2022). Khosravi et al. (2022) provide an overview of several systems with diverse objectives — giving feedback on writing, personalized hints in learning, assisting in providing peer feedback — which used XAI to explain their respective interventions. This means that the systems tried to explain to the users why they received a particular piece of information or advice. The intervention itself, however, is usually constructed manually. The conclusion drawn by Khosravi et al. (2022) is that the explanations were generally well received and that users stated a preference for systems with XAI; but also that there were some pitfalls which often relate to the still open question of what adequate explanations are in the context of interventions. Conati et al. (2021) similarly evaluated whether students preferred receiving explanations for hints in an intelligent tutoring system. Their results were generally positive, indicating that explanations increased trust and perceived usefulness. However, they found that whether and how someone preferred their explanations might depend on personal characteristics. While these studies mostly used pre-defined interventions and then explained them with XAI, Mu et al. (2020) predicted students in danger of wheel-spinning and then used the explanation for the prediction as interventions; i.e., when a feature was predicted to be the most important, they suggested changing the value of this feature. They found that the interventions based on these explanations generally matched what experts would have prescribed. This is in line with Hur et al. (2022)'s findings, who use SHAP to recommend interventions based on the predictions of students' performance. However, Mu et al. (2020) also make two important remarks: first, the explanations should refer to actionable features. If we suggest to a student in danger of wheel-spinning that they should review certain materials, that is good advice. But if we suggest, they should change their age, that is not helpful at all. Their second important remark is that SHAP explanations reveal nothing about causality. We believe that this, again, in part relates to contrastive explanations which LIME and SHAP do not provide. If features receive a different value, the person might still not receive a different prediction. But moreover, when using explanations as interventions, we do not know whether these relationships actually hold in real life. Imagine that a person in danger of wheel-spinning at a certain point in their learning process receives the advice to revise previous materials because the ML-model discovered a strong correlation between these variables, and the XAI-model picked up on this discovered relationship. Maybe the real connection between these variables is not that understanding the previous materials helps mastering the new content but simply that both these materials and the current content were covered in a previous course. In this scenario, revising the materials would not help much. This shows that the explanations XAI provide are limited and — as we have

Figure 3: An example of causal modeling explaining factors influencing whether Course 1 (C1) is dropped or not.



defined in the beginning — do not allow for interpretability.

4. A Model of Interpretability

To further illustrate our argument as well as our goal, which ultimately is to derive interventions and interpretability from ML-models, consider Figure 3. 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 complement the contents of C1, and taking them in parallel increases the competencies 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. However, all three classes, C2, C3, and C4, contribute to the workload. Having a high workload decreases the probability of finishing C1 dramatically. Imagine now that for a student, Bob, the drop-out probability for C1 is predicted to be high. What does that mean? Why is it high? How can he change his prediction? In order to fully leverage this prediction, we should walk through each of the three steps of interpretability that we have defined as follows:

1. *Understand which features matter.* This means we know which features matter for a person's prediction (as explained above). In our example, this means that the most important features regarding Bob's prediction are revealed to be the parallel taking of C2, C3, and C4. Furthermore, the XAI-methods at this step are likely to reveal that all three have a negative impact on completing C1. As we said before, this explanation is interesting but certainly not complete or even correct, considering the positive impact of C2 and C3 on competence and the positive of competence on completion. But even more, knowing the direction and impact of features is not equal to knowing what has to change in order to change the prediction. Should Bob 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 counterfactual 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 a basis for sensible advice. Note, though, that this does still not mean that the best advice is given. Maybe we find that not taking C2 changes the prediction but given the information from Figure 3, this is clearly not the best possible 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 different causal relationships 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 a correct and potentially the best intervention for Bob; moreover, we can use this knowledge to construct better programs for all students.

As mentioned before, if we use LIME and SHAP off-the-shelf, we can extract the most important features per person, but counterfactual explanations or causal connections are not provided. This limitation of XAI, particularly regarding counterfactual explanations, has been addressed and sometimes dealt with by other scholars as well (Keane and Smyth, 2020; Keane et al., 2021; Adadi and Berrada, 2018). We will argue, however, that XAI-methods provide a fruitful basis for causal analysis, which social scientists can also use to gain more insights. Nonetheless, XAI-methods can only provide a basis. How can we reach the other steps of interpretability? We argue that in particular reaching the third step calls for turning towards and employing techniques of social science. First, we need an understanding of the concept of counterfactuals and aim to extend LIME and SHAP in that way. Second, we need a theory-, instead of a data-driven approach to explore causal mechanisms. Social science is well equipped for this task and can, as we will discuss later, benefit from computer science and XAI in return.

4.1. Reaching Step 2: Counterfactuals

Reaching step 2 is easily possible as it rather naturally extends the idea of LIME (but is also readily applicable to SHAP) 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 (or less) features extracted by LIME and SHAP 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 — change the prediction and return the features and changed values[4]. 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. We always look for the smallest necessary change. 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 biggest change counts towards the selection of the top c features. For numerical features, we propose shifting the value a standard deviation

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

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 to change the prediction and how this change would need to look. When we advise students based on this, we have to ensure that we only report variables they can actively change.

4.2. 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 and SHAP return (positive and negative impact) as a basis for a deeper analysis. For this theory-driven approach, we propose to follow the steps:

Feature Extraction and Clustering. Extract all relevant features and their impacts and use them 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 (as already discussed in the XAI-section of the work).

Theorize about Causal Mechanisms. 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. This is very different from the typical proceedings when building a model. If you find that other features could also be important, can serve as a proxy-variable, or decompose the effect of a variable, add them. 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. Ideally, domain experts are consulted for this.

Model Causal Mechanisms. 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 (as in our example)? Does the latent or another variable mediate the effect? (b) Is there an actual relationship between two variables, or are they confounders, meaning that a third variable affects both? (c) Does a third variable moderate the effect between two variables? (d) Is the effect linear or quadratic? For computer scientists, in difference to social scientists, it is a little unusual to think about causal mechanisms, which is why the differences between these questions may not be obvious. A variable acts as a mediator if it is affected by another variable that also has a direct effect on the outcome and itself affects the outcome in the opposite direction. As an example, think about heart disease as an outcome. Smoking increases the probability of developing a heart condition, but smoking also decreases body weight, and a smaller body weight decreases the probability of developing a heart condition. A variable acts as a moderator if it impacts the relationship between another variable and the outcome. For example, if we want to see whether the difficulty of a class impacts how happy students are with the class, we may not find a statistically meaningful effect. But if we consider the motivation of the students, we may find upon using an interaction between motivation and difficulty that motivated student rate difficult classes better and easy classes worse and that this effect reverses for non-motivated students. Confounders are variables that seem to affect one another, but in reality, both are af-

fected by a third variable which should be added as well. A variable can also have a non-linear effect on the outcome. For example, students who are a little older may do better in classes at university as they are more focused. Much older students may have additional responsibilities such as childcare and may thus do worse. The effect of the variable would be quadratic, then. These differences should be clear when modeling the mechanisms.

Model a Regression Term. 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. When modeling the regression term, we again consider the different causal mechanisms leading us to construct the DAG. Each variable believed to have an effect is entered into the formula on its own. A moderator variable should not be entered on its own but in an interaction term with the variable actually having the effect. Mediators are added both on their own and in an interaction. Quadratic and other effects are modeled by having an additional term with the squared variable etc.

Evaluate the Results. Check the effects of the terms of your formula. What is significant and on what level ($alpha = 0.05, alpha = 0.01, alpha = 0.001$), i.e., is it likely that the variable actually impacts the outcome meaningfully, or could it be by chance? Does the direction conform to your theoretical considerations, i.e., does a variable thought to have a negative effect have a negative weight? Note that, in particular, when entering many variables, we run into the danger of getting a Type I error, meaning that we receive a significant result that is not actually significant (at $alpha = 0.5$, we get a Type II error in one out of twenty cases). Furthermore, the power of the test increases with many instances meaning that this reduces the Type II error of not finding an effect even though there is one but eventually — when we use really large datasets — we find significant effects that are not really meaningful. Thus, evaluation should be done very carefully. We can even check whether our causally-motivated regression formula does better than a formula without any interactions and quadratic effects by running two regressions and comparing the R^2 values. If the regression without interactions does not fare much worse, our causal ideas may not be correct.

Construct Intervention Systems. based on our results and combined with the insights of step 2, we can now construct personalized interventions and support/feedback. In particular, when constructing interventions, we should be careful when it comes to unwanted effects and design them carefully. We will discuss this again later.

5. EVALUATION

In order to demonstrate the pipeline and its applicability, we test our approach on an artificial and a real-life set of data. The datasets will be described first; then, we will compare SHAP and LIME before evaluating the remaining steps.

5.1. DATA

Artificial Data. As we typically do not know the real causal mechanisms and can only make informed guesses considering the existing literature and our own reasoning, we test our pipeline on artificial data. Regarding the counterfactuals, we can test whether it returns the feature sub-

set 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[5]. 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 (+) behind the variable indicates a positive effect on drop-out, and a minus (-) indicates a negative effect. The number of symbols (e.g., ++) represents the strength of the effect as each causal relationship was given a weight by which the probability of a drop-out changes. For example, $V6$ has a stronger effect than $V7$. Features with no effect have a "No" behind them. Features having an effect in combination with other features are connected with a plus sign. The effect of their combination is presented after the colon. We created 10,000 instances by randomly sampling features. The drop-out value was determined by the sum 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's degree at the University of Mannheim, Germany.[6] 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 more than once. 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 includes 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 course A. This course has four features assigned to it: whether it 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 when the student registers for the course we are predicting. In total, we consider 30 courses. Table 2 provides summary statistics of the data. "Year" refers to the study year a person is currently in; e.g., 1 would mean that a person is in their first year of studies. By domestic nationality, we mean German nationality. Likewise, a domestic High School degree is a German high school degree.

5.2. STEP 1: LIME AND SHAP

For both data sets, we predicted 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 an F1-value of 0.99. For

[5]The artificial dataset as well as some additional files illustrating how to continue with the next steps can be found here: https://github.com/lea-cohausz/jedm.git.

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

Table 1: Artificial Data — variables, their effects, and what could be recovered using LIME and SHAP 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(No) + V18(No): ++	V19, V20, V21, V22, V23, V24, V25, V26 (No effect)
Recovered (LIME)	V1(+), V2(+), V3(+), V4(+), V5(+)	V6(−), V7(−), V9(−), V10(−)	V11(+), V12(+), V13(+), V14(−),V15(−), V16(−)	V20(−)
Recovered (SHAP)	V1(+), V2(+), V3(+), V4(+), V5(+)	V6(−), V7(−), V8(−), V9(−), V10(−)	V11(+), V12(+), V13(+), V14(−),V15(−), V16(−)	V19(−/+), V20(−), V21(−/+), V22(−), V23(−), V24(+), V26(−)
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(No): +; V14(+) + V15(+) + V16(No): −; V17(No) + V18(No): +	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%

the real-life 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 (or less, if SHAP did not assign importance to that many variables) and their directions for each instance of the test data. We only considered those instances for which drop-out was predicted (318 in the artificial data and 26 in the real data) as these are the ones we are most interested in.

LIME vs. SHAP. Although LIME has often been criticized for providing less meaningful and consequent explanations due to its sample-based approach, it appears to be much better than SHAP at recovering only those features that matter according to our data generation procedure. In fact, when looking at what features got extracted at least for one instance, we can see that SHAP extracted all features but $V25$ (and sometimes multiple times), whereas LIME generally only extracted the features we intended to have an effect when designing the dataset. In particular, when interactions were at play (e.g., when an instance had a 1 for variables V11-V13 which turns around the effect of V11 and V12), SHAP tended to assign importance to a random set of features, whereas LIME extracted all relevant features. To provide an example for this:

We noticed a pattern in SHAP that when the variables $V11$, $V12$, $V13$ all had value 1 assigned, meaning that the effect of all three is positive although the individual effect of $V11$ and $V12$ is still negative, SHAP would assign a positive effect to $V11$ and $V12$ and a negative to a variable not having an effect according to our dataset design, e.g., $V19$. The most often extracted features and their assigned effect sizes and directions were the same for LIME and SHAP, though. Therefore, we now continue with the results for LIME only.

Artificial Data. Table 1 shows which features were 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. Again, consider as an example $V11$, $V12$, and $V13$. On their own, $V11$ and $V12$ have a negative effect on drop-out, but when all three variables have a 1 assigned to them, they together have a large positive effect - although the individual negative effects still persist. Both SHAP and LIME tend to attribute the positive impact to each of the variables involved in this ($V11$, $V12$, and $V13$). This shows the limitations of LIME (and also SHAP) and, therefore, the importance of our remaining steps.

Real Data. 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 (variable "Date") or in the resit (26)[7], 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).

5.3. Step 2: Counterfactuals

Artificial Data. We now selected those important features that positively affect 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, only one feature 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 and 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 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

[7]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.

one feature to change. This rose to 19 when we considered changes in features referring to the same course as just one feature. Again, remember that each course has four features assigned to it (whether it is taken in parallel to C1, whether the student failed it, whether the student dropped out, or whether the student passed). There were ten instances for which only one specific feature changed the prediction but not any of the others. For the remaining four instances more than one feature had the ability to — on their own — change the prediction. Two instances needed two changes, and 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 related to two of the three courses extracted as important before (16, 13), the variable indicating the date (16), the age (8), and the 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.

5.4. Step 3: Regression and Modelling

Artificial Data. For the artificial data, we simply used all the extracted features and our knowledge [8] 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 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.

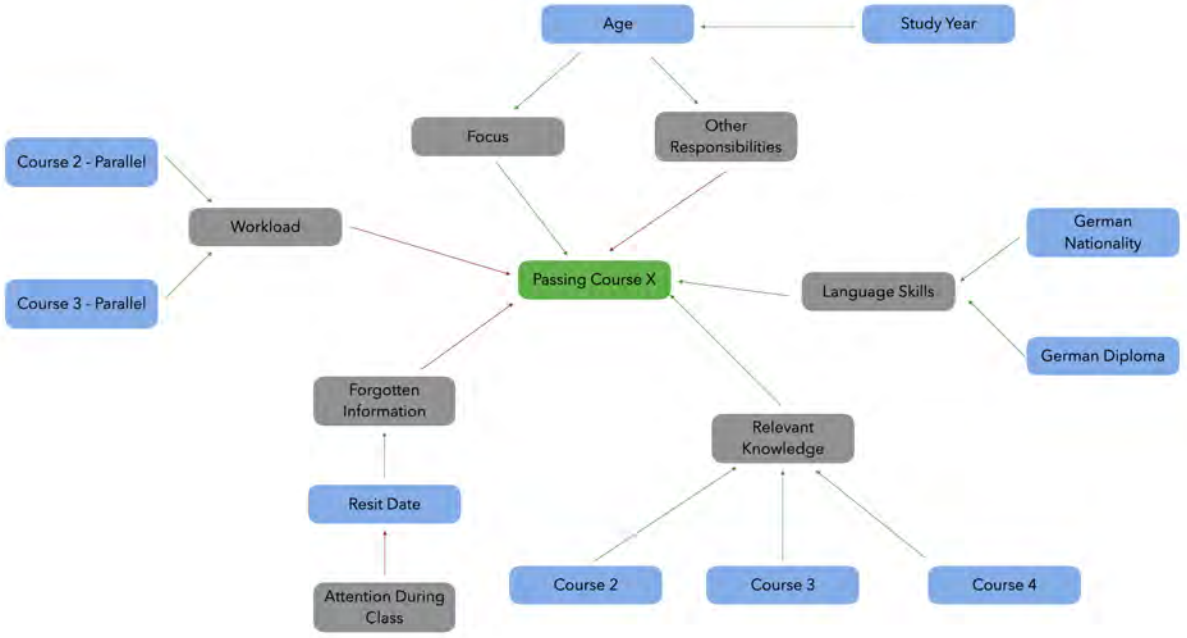
Real-Life Data. For the real-life data, as explained above, we first clustered our test instances using k-Nearest Neighbor based on the features extracted so that we only focus on features relevant to this set of students. We chose $k = 3$ upon visual inspection. For illustrative and space reasons, we will focus on the largest cluster containing twelve instances. We used all features that were extracted for more than one instance in the cluster (Table 3).

We will now discuss what we theorize regarding these features and how — based on the theories — we construct the DAG[9] and the corresponding regression terms. The resulting DAG can be found in Figure 4. Note that variables with a grey background are variables that we do not have data on but that we believe to be influential constructs and that are connected to variables that we do have data on. The colors of the arrows show whether a variable has a positive (green) or negative (red) impact on the other variable it influences.

[8] Hence, also entering $V17$ and $V18$ again.

[9] For constructing DAGs and also for then estimating specific effects of variables based on the information on dependencies and independencies found in the DAG, we can advocate using DAGitty, a browser-based modeling tool which also has an easy to use R package (Textor et al., 2016).

Figure 4: The DAG of our real-life example.



The only two features that are not specific to our setting are the variables age and nationality (the domestic/German nationality variable mentioned before). 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 (Spitzer, 2000; Yasmin, 2013; Chen, 2012). 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 (older than 30) by the latter; this is represented by the DAG and, 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 for the negative effect LIME implies is that the degree is in German, which creates a language barrier for non-German speaking students (Morrice, 2013; Evans and Morrison, 2011). This might be mitigated when a student has already received their high school diploma in Germany. The DAG shows how the two variables interact, and furthermore, 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 (variable "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 not have paid enough attention to this course during the lecture period and b) students may have forgotten important information in the meantime.

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, $--$
$passC2 * passC3 * passC4$		$-$

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 which is also modeled by the DAG.

Table 3 shows our results; the middle column summarizes the results of the logistic regression without interaction or quadratic terms, and the right-hand column the results, including these terms. Again, 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. However, 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 course also have a higher probability of dropping C1; passing C3 also means that C1 will likely be successfully 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.

5.5. What could we do with that?

What do we take away from this? The last step of the pipeline consists of now using the information we gained through the pipeline. In our case, regarding the demographic features, we should investigate whether the age effect — that comparatively very old and young students struggle in difference to slightly older students — persists across the overall program. If so, we

should think about how to help these age groups (of course, telling people that their age may be a hindering factor is something we should not do). In order to do that, we could a) consult social science literature again to understand why these groups struggle and b) conduct surveys asking the students themselves. In that way, we can better understand the mechanisms at work and then use appropriate methods to intervene. Furthermore, considering the university-specific features and the predicted course, in particular, we should encourage the students not to choose the resit date. We should also identify students who have struggled with courses of similar content before and offer increased assistance and attention to them. How to best do this is, again, something social science may also help us with. But, as we will now briefly discuss, designing such interventions needs to be done carefully.

6. Alice, Bob, and the Potential Effects of Knowing Your Prediction

6.1. Students' Perspective

At the beginning of the work, we introduced Alice, who has a high probability of finishing a class, and Bob, who has a low probability of completing the same class. Initially, we said that this is not enough to do anything but when we employ our pipeline, we may now have a lot more information and can provide them and, in particular, Bob, with some advice. Again, if we find effects of variables they cannot change (as their age), obviously, we cannot mention this to them. But what about other changeable aspects? We argue that all information about predictions in this area have to be handled very carefully. The first question is not: How do we design the advice and intervention? But: Do we even tell them of their prediction? Indeed, telling them about the prediction is already problematic. Bob may be discouraged and demotivated, and Alice may feel too confident and, in response, do too little. We believe that telling people their exact prediction is generally not a good idea and that, at the very least, there should be an opt-out option to learn the prediction. Instead, students should ideally receive advice and interventions tailored to them but do not lead them to believe they are doing bad (or very good) anyway.

As already mentioned, Khosravi et al. (2022) and Conati et al. (2021) evaluated how students perceived receiving explanations along with their interventions. The general finding seems to be that most people prefer seeing an explanation. In our framework, intervention and explanation could rather naturally go hand-in-hand due to our causal understanding. In difference to, e.g., Conati et al. (2021), we probably do not have too big of a problem with explanations being perceived as annoying, as we can combine intervention and explanation. Furthermore, another difference would be that Conati et al. (2021) explain to people why they receive the intervention by usually stating that the system perceived them to have a problem. We, however, and a bit more similar to some of the systems evaluated by Khosravi et al. (2022), would not state that they receive the intervention because the system thinks they have a problem, i.e., might be in danger of dropping out or failing. For example, if a student decides to sign up for a course and the predictive model predicts the student to drop out, and we learn that this is likely because they have not yet taken another course closely related in content; then the intervention should be that the student should consider first taking the other course and the explanation along with it should be that this course teaches contents relevant for the course the student originally wanted to take. Again, intervention and explanation go hand-in-hand. Some other example interventions may

be: Before the semester starts, we could advise students on their workload or classes they should also take based on their data regarding their previous courses. As the semester continues, we can provide students with (partially) individualized reminder emails to revise certain topics and with special exercises, etc. If we have activity data and uncover that a student is likely to drop out soon, we can try to reach out to them. Again, we should always provide an explanation along with it as evidence points to this increasing trust and willingness to listen (Khosravi et al., 2022; Conati et al., 2021) and because our framework allows us to seamlessly do it.

However, note that there is an obvious danger in this: extra reminders and emails may annoy students or make them feel like the workload is too high, leading to the exact opposite effect. Furthermore, students may feel belittled when we advise them on their choices — again, the explanations might be able to help. We want to once again stress that it is very important to be careful and that psychological research can probably provide relevant insights for designing such systems. Furthermore, the introduction of such systems should be accompanied by a careful scientific evaluation recording the effects.

6.2. Instructors' Perspective

Generally, we believe that instructors can also benefit from our framework and receive advice along with explanations with the advice. Again, though, this has to be handled carefully. For example, instructors should probably not receive information on who exactly is likely to drop out or fail as this might result in instructors not investing as much in these students — leading to a confirmation bias Oswald and Grosjean (2012). Instead, they should maybe receive hints as to which student may need help with which materials. For our evaluation example, an instructor may learn that some students have not taken a relevant course before and that they should offer additional materials on this course. If we have activity data, we could alert instructors regarding what materials students generally struggle with.

6.3. Administrators/Managements' Perspective

Information provided to the administrators or study management might be less problematic regarding the psychological effects. Generally, either information on individual students or information on programs or courses as a whole can be passed on along with an explanation. An example of the former: If a student is predicted to drop-out of their study program because they have struggled with completing a few courses, then this can be made known to a study manager who can contact the student. An example of the latter might be that we find older students to struggle due to scheduling issues. The administration can then decide whether they can facilitate the participation of these students, e.g., by providing child care on campus or allowing for more flexible timetables.

7. Combining the Power of Two Disciplines

Before we conclude this paper, we want to briefly consider the topic of collaborations between social scientists and computer scientists. In this paper, we have repeatedly mentioned that social science plays an important role in this topic. DAGs and causal mechanisms belong to social scientists' daily tools and topics, whereas computer scientists are not typically concerned with this. Causal modeling, the third step in our pipeline, requires theory and past empirical work

from the social sciences — also not something computer scientists typically work with — and when constructing advice and interventions, we should again consult, in particular, psychological research to do that. The whole paper, thus, shows that we think that EDM as a whole can greatly benefit from collaborating with social scientists but we also think that social science can greatly benefit from computer science. The mindsets of the two disciplines are very different, which can be an initial hurdle but eventually also prove an advantage.

While computer scientists are typically focused on accuracy values and comparing their predictions, social scientists do not compare their predictions, as they are rarely ever focused on them and instead try to understand mechanisms. When it comes to understanding the models and the world beyond probabilities, computer scientists may thus be a little at a loss, and this could also be why there are so few studies in EDM focusing on explaining models. Computer scientists, however, have the ability to explore the data and simply use those features that the model and not theory shows them to be important. They are, therefore, able to uncover interesting new mechanisms, whereas social scientists tend to stay with the theory. Working together with computer scientists may, thus, allow them to uncover new causal mechanisms. Furthermore, the focus on applications (and, as a result, necessarily predictions) is unusual for social scientists but can also benefit them and provide further new insights. Hence, we firmly believe that the two disciplines should try to work together more often, even if it takes some time to understand each other.

8. Conclusions, Limitations, & Future Work

In this paper, we argued that only knowing predictions of student success and drop-out in the form of probabilities is not enough when we aim to use these models for personalized interventions. Using XAI-methods, such as SHAP and LIME, can provide an explanation for the model's decision and a basis for interpreting the world with the model, but are, on their own, also not enough. Therefore, we attempted to provide a framework enabling us to use the information extractable from predictions as a basis for a personalized intervention system and to provide feedback and support to instructors and administrators. To highlight the challenges and requirements, we came up with a step-wise model of interpretability where step 1 means identifying important features (which can be done with classical XAI-methods), 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 their applicability. We also concluded that using LIME achieves more reliable results than using SHAP. Our results on artificial data showed that the method works well when we correctly theorize about causal mechanisms. Of course, we may not always be 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 and our discussion of XAI-methods 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 endeavor. 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 argued for 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, usually do not work on predictions but are knowledgeable regarding statistical tools to uncover causal mechanisms and derive models from theory. While these differences in approaches are prone to hinder collaboration, this task 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. Thus, we believe that future work should focus on combining the two disciplines to receive better models and explanations.

9. ACKNOWLEDGEMENTS

I want to thank Sarah Alturki, who provided great feedback on short notice and, I believe, helped improve the work a lot at an early stage. Furthermore, I am grateful for the immensely helpful and kind reviews provided by the reviewers of JEDM.

REFERENCES

ADADI, A. AND BERRADA, M. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access 6*, 52138–52160.

ALAMRI, R. AND ALHARBI, B. 2021. Explainable student performance prediction models: a systematic review. *IEEE Access 9*, 33132–33143.

BARANYI, M., NAGY, M., AND MOLONTAY, R. 2020. Interpretable deep learning for university dropout prediction. In *Proceedings of the 21st Annual Conference on Information Technology Education*. Association for Computing Machinery, 13–19.

CHEN, R. 2012. Institutional characteristics and college student dropout risks: A multilevel event history analysis. *Research in Higher Education 53,* 5, 487–505.

CHITTI, M., CHITTI, P., AND JAYABALAN, M. 2020. Need for interpretable student performance prediction. In *2020 13th International Conference on Developments in eSystems Engineering (DeSE)*. IEEE, 269–272.

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

CONATI, C., BARRAL, O., PUTNAM, V., AND RIEGER, L. 2021. Toward personalized xai: A case study in intelligent tutoring systems. *Artificial Intelligence 298*, 103503.

DEL BONIFRO, F., GABBRIELLI, M., LISANTI, G., AND ZINGARO, S. P. 2020. Student dropout prediction. In *International Conference on Artificial Intelligence in Education*, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Springer, 129–140.

EMMERT-STREIB, F., YLI-HARJA, O., AND DEHMER, M. 2020. Explainable artificial intelligence and machine learning: A reality rooted perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10,* 6, e1368.

EVANS, S. AND MORRISON, B. 2011. Meeting the challenges of english-medium higher education: The first-year experience in hong kong. *English for Specific Purposes 30,* 3, 198–208.

HUR, P., LEE, H., BHAT, S., AND BOSCH, N. 2022. Using machine learning explainability methods to personalize interventions for students. In *Proceedings of the 15th International Conference on Educational Data Mining*, A. Mitrovic and N. Bosch, Eds. International Educational Data Mining Society, 438–445.

KEANE, M. T., KENNY, E. M., DELANEY, E., AND SMYTH, B. 2021. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence Organization, 4466–4474. Survey Track.

KEANE, M. T. AND SMYTH, B. 2020. 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*, I. Watson and R. Weber, Eds. Springer, 163–178.

KHOSRAVI, H., SHUM, S. B., CHEN, G., CONATI, C., TSAI, Y.-S., KAY, J., KNIGHT, S., MARTINEZ-MALDONADO, R., SADIQ, S., AND GAŠEVIĆ, D. 2022. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence 3*, 100074.

LANGER, M., OSTER, D., SPEITH, T., HERMANNS, H., KÄSTNER, L., SCHMIDT, E., SESING, A., AND BAUM, K. 2021. What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence 296*, 103473.

LUNDBERG, S. M. AND LEE, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus, Eds. Vol. 30. Curran Associates Inc., 4768–4777.

MANRIQUE, R., NUNES, B. P., MARINO, O., CASANOVA, M. A., AND NURMIKKO-FULLER, T. 2019. 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*. Association for Computing Machinery, 401–410.

MESKE, C., BUNDE, E., SCHNEIDER, J., AND GERSCH, M. 2022. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management 39,* 1, 53–63.

MILLER, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence 267*, 1–38.

MORRICE, L. 2013. Refugees in higher education: Boundaries of belonging and recognition, stigma and exclusion. *International Journal of Lifelong Education 32,* 5, 652–668.

MU, T., JETTEN, A., AND BRUNSKILL, E. 2020. Towards suggesting actionable interventions for wheel-spinning students. In *Proceedings of the 13th International Conference on Educational Data Mining*, A. N. Rafferty, J. Whitehill, C. Romero, and V. Cavalli-Sforza, Eds. International Educational Data Mining Society, 183–193.

OSWALD, M. E. AND GROSJEAN, S. 2012. Confirmation bias. In *Cognitive Illusions*, R. Pohl, Ed. Psychology Press, 91–108.

PRENKAJ, B., VELARDI, P., STILO, G., DISTANTE, D., AND FARALLI, S. 2020. A survey of machine learning approaches for student dropout prediction in online courses. *ACM Computing Surveys (CSUR) 53,* 3, 1–34.

QIU, L., LIU, Y., HU, Q., AND LIU, Y. 2019. Student dropout prediction in massive open online courses by convolutional neural networks. *Soft Computing 23,* 20, 10287–10301.

RIBEIRO, M. T., SINGH, S., AND GUESTRIN, C. 2016. Model-agnostic interpretability of machine learning. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, B. Kim, D. M. Malioutov, and K. R. Varshney, Eds. arXiv:1606.05386.

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

TEXTOR, J., VAN DER ZANDER, B., GILTHORPE, M. S., LIŚKIEWICZ, M., AND ELLISON, G. T. 2016. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International Journal of Epidemiology 45,* 6, 1887–1894.

XING, W. AND DU, D. 2019. Dropout prediction in moocs: Using deep learning for personalized intervention. *Journal of Educational Computing Research 57,* 3, 547–570.

YASMIN, D. 2013. Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. *Distance Education 34,* 2, 218–231.

ZEINEDDINE, H., BRAENDLE, U., AND FARAH, A. 2021. Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering 89*, 106903.