PennState

Factors Impacting K-12 Teachers in Understanding Explanations of Machine Learning Model on Students' Performance

BY

Hangzhi Guo

Na Li

December 15, 2020

College of Information Science and Technology

Pennsylvania State University

Contents

Figures

List of Tables

Factors Impacting K-12 Teachers in Understanding Explanations of Machine Learning Model on Students' Performance

Hangzhi Guo,[1] **Na Li,**[1]
[1]Pennsylvania State University, University Park, PA 16802
{hangz, nzl5264}@psu.edu

December 15, 2020

Abstract

In 21th century, artificial intelligence (AI) is pervasive in our life and decision-making process. It is increasingly important to determine whether humans can trust AI's decisions or not. To gain humans' trust in AI, making AI's decisions more interpretable to humans is one of the key approaches. Although explainable AI is an actively researched field, most efforts are put into enhancing the interpretability for data scientists. Many methods invented to make the AI's decision more interpretable heavily rely on statistical visualizations. Such a level of interpretability to humans, while possibly sufficient for ML experts, might not be sufficient for people without an ML background to understand. In this user study, we investigated both teachers and ML practitioners' use of the interpretable AI model to predict students' final scores. We aim to study the differences of users' backgrounds on how they interpret the AI model. We generated design implications to improve the current explainable AI.

***K*eywords** XAI · Interpretability · Machine Learning · User Study · K-12 teachers · Human-Computer Interaction

1 Introduction

Machine learning (ML) is pervasive in our daily lives and it has been applied in different disciplines to help people make decisions. However, the machine learning model such as the random forests and deep neural networks is notoriously difficult for ML practitioners to understand (Kaur et al., 2020), not to mention the lay person who does not have data science or machine learning backgrounds. In face of the challenges, machine learning practitioners try to use some techniques to present the ML model in a way to make it more understandable to people even though they are not expertise in machine learning, which is called explainable AI (XAI) or interpretable ML (Lipton, 2018; Gilpin et al., 2018; Murdoch et al., 2019). The interpretability refers to the extent to which a system explanation is understandable to humans. Although the interpretable ML is gaining increasing popularity in the machine learning community, the factors impacting human's understanding of it have been rarely investigated.

In this study, we developed an explainable machine learning system which not only predicts the students' final scores, but also presents the explanations of such predictions in an online dashboard. We presented three different explanations to the users: feature-based explanation, rule-based explanation and counterfactual-based explanation to help them predict students' final scores. This pilot study is aimed at analyzing the underlying reasons responsible for human's understanding of the ML interpretability tools. We focus on two stakeholder groups: teachers with K-12 teaching experience but without machine learning background, and machine learning practitioners without teaching experience.

Our user study consists of three phases: 1) delivered surveys ($N = 4$) to identify the backgrounds of the users such as whether they have teaching or machine learning experience. 2) video-recorded the process about how the users use the model ($N = 4$) and 3) follow-up interviews to understand the issues and difficulties the users encountered when using the model.

Our results indicate that all of the users prefer using the feature-based explanation to predict students' final scores because they thought it is more straightforward to understand than the other two. Furthermore, we found the prior experiences of the users exerted considerable influence on their understanding of the model. For example, the predictions of the teachers' on students' final scores are different from the machine learning practitioners because the former ones prefer using their teaching experience to judge the scores of the students, while the latter made their predictions based on knowledge of machine learning models. Overall, our results highlight the differences

between teachers and ML practitioners in their understanding of the model as well as the challenges they confronted in the process.

2 Related Work

2.1 Interpretability in Machine Learning

To accommodate the requirement of General Data Protection Regulation (GDPR) to provide "meaningful information about the logic involved" by algorithmic methods (Voigt and Von dem Bussche, 2017), there is a surging trend in the machine learning community to ensure the ML models' predictions being interpretable to humans. There are two main approaches to accomplish this goal: applying "glass-box" models that are intrinsically interpretable, and explaining the "black-box" models in a post-hoc manner.

Rudin (2019) argues the advantages of using interpretable models over complicated black-box models. First, the explanations from the interpretable models are faithful to the model itself, while it is likely not the case for black-box models. Second, the explanation from the interpretable models make a complete explanation, while post-hoc explanations can only provide partial explanations of the black-box models. Third, interpretable models do not indicate compromises over the predictive performance; in fact, two types of models, *generalized additive models* (Lou et al., 2013; Caruana et al., 2015) and *rule-based models* (Letham et al., 2015; Lakkaraju et al., 2016), have been successfully applied to support high-stake decision making in real-world problems. Motivated by the rule-based models, we investigate how IF-THEN explanations are interpretable to users in this study.

Although researchers enjoy the utter transparency of the interpretable models, some complicated real-world problems, such as image classifications, sentiment analysis, cannot be successfully modeled by simple interpretable models. Therefore, extra efforts need to be taken to reveal the mythos behind those models. To explain the black-box models, most approaches work as a post-hoc manner (after making the prediction). Generally speaking, the current explanation methods of post-hoc analysis can be categorized into three tracks, *feature-based explanation*, *counterfactual explanation*, and *case-based explanation*.

The *feature-based explanation* is perhaps the most frequently used and investigated approach in interpreting a predictive model. It measures how important a particular feature of the predictive model. Ribeiro et al. (2016) propose Local Interpretable Model-agnostic Explanations (LIME), which uses a linear model to provide local explanation of any black-box model. LIME only explains how the model predicts one particular data instance. It samples data near the explainee (data instance to be explained), and uses a linear model to fit those data. Similarly, Lundberg and Lee (2017) propose SHAP

(SHapley Additive exPlanations). SHAP also measures the feature importance locally, but it calculates the Shapley value defined in game theory. In our research, we apply SHAP to generate the feature-based explanation.

The counterfactual explanation offers a contrastive case that provides an opposite outcome with changes in the input. For example, if a person applies for a loan and gets rejected, he/she might be more interested in what it would take to get a loan. In other words, the counterfactual explanation answers the what if questions. Wachter et al. (2017) proposed a method to generate the counterfactual explanation by minimizing the distance between input instance (to be explained) and the counterfactual example, and pushing the new prediction toward the desired classes (e.g. pushing the model's prediction from getting rejected to getting approved). Other algorithms, built on top of Wachter et al. (2017)'s method, optimize other aspects, such as diversity (Mothilal et al., 2020), closeness to the data manifold (Van Looveren and Klaise, 2019), and causal constraints (Mahajan et al., 2019). In our research, we employ Wachter et al. (2017)'s method to generate the counterfactual explanations.

The case-based explanation (also called example-based explanation) finds the most similar cases in the dataset. It is motivated from the human reasoning process on "reason through analogy" (Wang et al., 2019). For example, a doctor might diagnose a disease to a patient because another patient with similar symptoms ended up with that disease. Some models in ML, such as clustering, collaborative filtering, K-nearest neighbors, are motivated by such a case-based reasoning process. However, most black-box models, such as neural networks and ensemble models, are not case-based learning methods. Caruana et al. (1999) first identified such difficulty by proposing a case-based explanation over neural networks via computing the Euclidean distance between the input instance and all training samples in its latent space. More recently, Chen et al. (2019) propose a novel explanation style, "this looks like that", to explain image classifications by identifying the similar images and their similar regions. Kanehira and Harada (2019) propose a neural network architecture to generate a complementary explanation, which searches for both similar and alien cases. Due to the time constraints, our current study does not implement this explanation style. We plan to further investigate this in the future research.

2.2 HCI Research in Interpretability

HCI research focuses on user centric study and analysis on evaluating how understandable the machine explanations are to humans.

First, it is imperative to identify the stakeholders of explainable AI. Ribera and Lapedriza (2019) identified three groups of users of the explainable machine learning: (i) Machine learning practitioners (e.g. AI researchers, data scientists) are a group of technical users who work with models and data who possess knowledge in machine learning and statistics. Their primary demand for explainable ML is to diagnose, debug and improve their models. (ii) Domain experts (e.g. doctors, K-12 teachers) are professions who have specialized expertise in specific domains. Typically, they do not have prior knowledge in machine learning (some of them might have statistics and a college mathematics background). However, many of them rely on intelligent systems to support their decision making process, and it is essential for them to know when to trust the model's prediction, and when not to. (iii) Lay person (e.g. applicants to loan) assumes people with any background. They need explanation for recourse. In our study, we focus on investigating machine learning practitioners and domain experts (K-12 teachers).

Many HCI studies try to identify the need of users in machine explanations. Miller (2019) analyzes literature in social sciences. He identifies four important attributes of an interpretable explanation: (i) explanations should be contrastive; (ii) explanations should be selected to aid the human decision; (iii) probabilities might not matter; and (iv) explanations are social and communicative. Wang et al. (2019) take one step further by connecting theories in social science to practices in machine learning. For example, they identify analogical reasoning as an essential process in human decision-making, which motivates case-based reasoning in machine learning. Moreover, they also analyze how humans suffer from biased decisions and propose strategies for mitigating human errors.

Finally, a lot of human studies and evaluations are conducted in the HCI community. Binns et al. (2018) show that the explanation style impacts justice perception. They design a user study to make criminal cases justice assisted with intelligent systems with four different explanation styles: input influence, sensitivity, case-based, and demographic. Their qualitative results suggest that people consider fairness questions in receiving the explanations of ML models. Lage et al. (2019) investigate the influence of explanation complexity to human interpretability using the interpretable decision set. Their results indicate the model complexity contributes little to users in making the correct decision, while the user satisfaction drops significantly. Kaur et al. (2020) conduct user-centrics evaluation over data scientists. Their tasks require data scientists to use the interpretable tools (such as LIME, SHAP) to diagnose the issues of black-box models. Their results show that all of the participants (including senior developers) make conceptual mistakes while using interpretable tools.

Table 1: Features used in the study and their descriptions (Cortez and Silva, 2008).

Feature name	Describe
G2	Second period grade (from A to F)
G1	First period grade (from A to F)
failures	Number of past class failures (numeric: n if 1<=n<else 4)
higher	Wants to take higher education (binary: 1 for yes or 0 no)
age	Student's age (numeric: from 15 to 22)
school	Student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
goout	Going out with friends (numeric: from 1 - very low to 5 - very high)
Mjob	Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
health	Current health status (numeric: from 1 - very bad to 5 - very good)
freetime	Free time after school (numeric: from 1 - very low to 5 - very high)
absences	Number of school absences (numeric: from 0 to 93)
Walc	Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
famrel	Quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
Medu	Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu	Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

3 Methods

3.1 Research Design

The goal of our research is to study the two different groups' use of the machine learning interpretability tools, and how their different backgrounds influence their understanding of the model. By using a qualitative method, we conducted pilot interviews with the users. The interview protocol was designed to understand the challenges users have in their process of understanding model explanations, and how they made predictions of students' final scores based on the model or their personal experience. On average, each interview lasted for 30 minutes. Through conducting inductive thematic analysis of the interview transcripts, we identified some themes capturing the common issues both groups had, and the factors impacting their understanding of the model.

Table 2: An overview of the four users' demographic information.

User	Highest degree	Prior experience	Data analysis experience
A	Master in TESOL	More than 3 years' K-12 teaching experience	No experience
B	Master in Education Policy	More than 2 years' K-12 teaching experience	Yes, used excel, STATA, R, Python in the study and work experience.
C	Bachelor in Computer Science	ML experience in a research project	Used python and excel to train some machine learning models.
D	Bachelor in Computer Science	ML experience in study	R and Basic level use some packages (ggplot) to get the graph.

3.2 Participants

We invited four participants to use our model and divided them into two groups (teaching group and ML group), each group with two people. One group had K-12 teaching experience and graduated with education master degrees, but they had no knowledge of machine learning. The other group graduated with computer science bachelor degrees and had machine learning experience, but no teaching experience. All of the participants are originally from China and they are fluent in English. There are minor differences in the teaching group; one teacher had prior experience in using STATA and Python for data analysis while the other teacher had zero experience in data analysis tools.

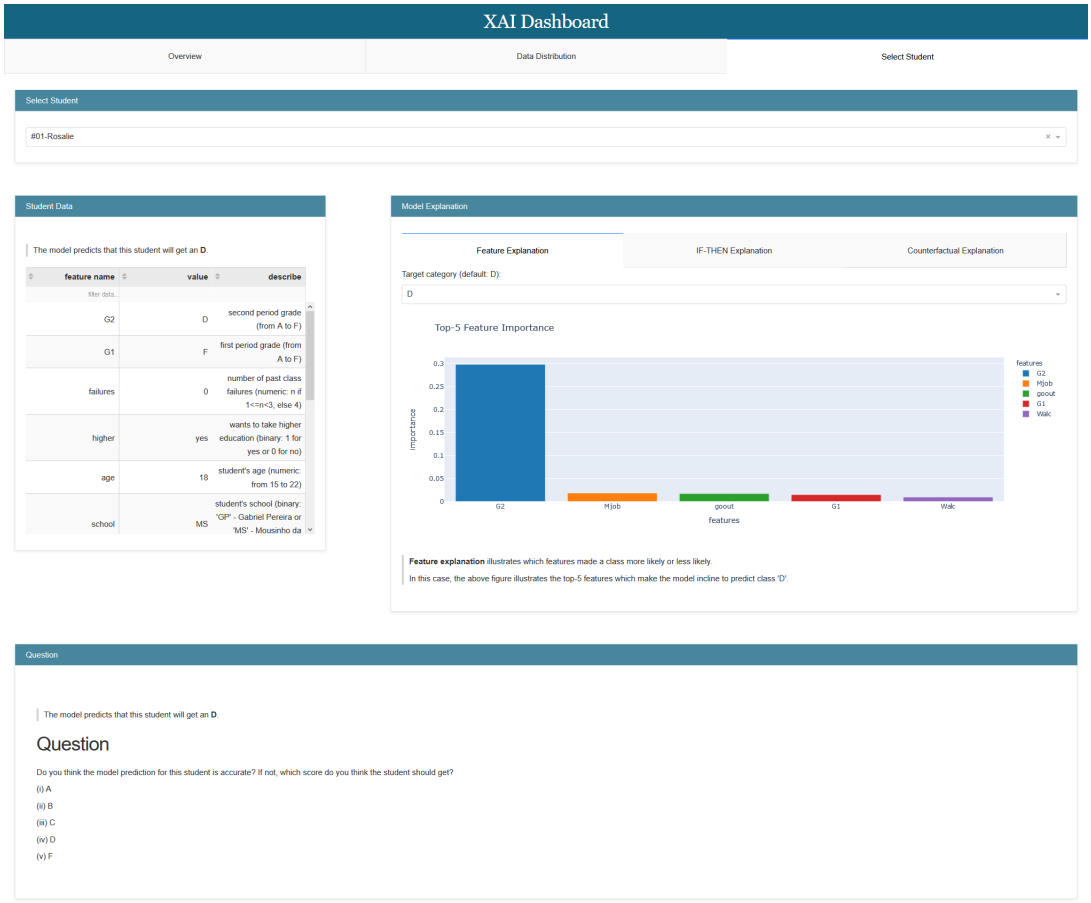

Figure 1: Screenshot of the online dashboard showing basic student information (middle left under the title *Student Data*), model's prediction (middle left and the bottom), and three explanations (middle right).

3.3 Overview of the Procedure

Before starting the user study, we developed an explainable ML interface for presenting the model explanations. We used a student performance dataset from the UCI machine learning repository (Cortez and Silva, 2008) to train a machine learning model to predict the students' final scores. First, out of 32 features in the original dataset, we manually select 16 relevant features (see Table 1). Then, we split the dataset into training and testing sets, each having 600 and 49 samples, respectively. Next, we choose the random forest (Breiman, 2001) to fit on the training set, and we select the parameter by cross validating on the training set. In the end, we train the model using the Scikit-Learn package (Pedregosa et al., 2011), and we set the estimator number equals to 150. All student data used in the user study will not be trained for fitting the ML models. Then, we will use state-of-the-art interpretable tools and models to explain the model prediction.

Specifically, we will use SHAP (Lundberg and Lee, 2017) for feature-based explanation, Anchor (Ribeiro et al., 2018) for rule-based explanation, and Wachter et al. (2017)'s implementation for counterfactual explanation.

Moreover, we created an online dashboard interface to display basic student attributes, the student's prediction result, and three different explanation styles (see Figure 1 below). We use Plotly Dash[1] to create the front-end page and data visualization, and Flask[2] for back-end serving. The dashboard app is deployed on Heroku[3]. This dashboard includes three interfaces: *Overview*, *Data distribution* and *Select Student*. In the *Overview* page, there is a brief description about what the users are supposed to do by using the model. The *Data distribution* page presents histograms for different features to gain insights of the shape of the data. The *Select Student* page includes performance data of 10 different students, and users can pick one student at a time to view his/her performance data. Besides, on this page, we present 3 different explanations: 1) feature-based explanation (Figure 2) which includes 5 top important features and their weights in determining a student final score, 2) if-then explanation (Figure 3) which explains individual predictions of the model by finding a decision rule that is sufficiently to make a prediction, in other words, changing other features will not alter the score, 3) counterfactual explanation (Figure 4) which describes a causal situation in the form "if feature X has changed, Y would also change". It describes how to change the minimum features to find a different prediction.

We invited four participants to use our model. After obtaining their consents, we delivered a survey (see appendix) to obtain basic information about their backgrounds including their highest degrees, working experience in teaching or machine learning, data analysis experience, and mathematics experience. Then, each of the participants was given 30 minutes to use the model and answered 10 questions about their predictions on the students' final scores. After that, we conducted follow-up interviews with them (see table 3) about the issues they encountered in using the model.

To find out the factors impacting the two different groups (Teachers vs. ML practitioners) understanding of the model, as well as the issues they encountered in making predictions based on the ML model. We performed an inductive thematic analysis on the interview transcripts of the four users. Two researchers participated in the process to do the data analysis. We held weekly meetings over two weeks to discuss the general findings generated from the interview data. Two major themes surfacing are the **common**

[1] `https://dash.plotly.com/`
[2] `https://flask.palletsprojects.com/en/1.1.x/`
[3] `https://dash-xai.herokuapp.com/`

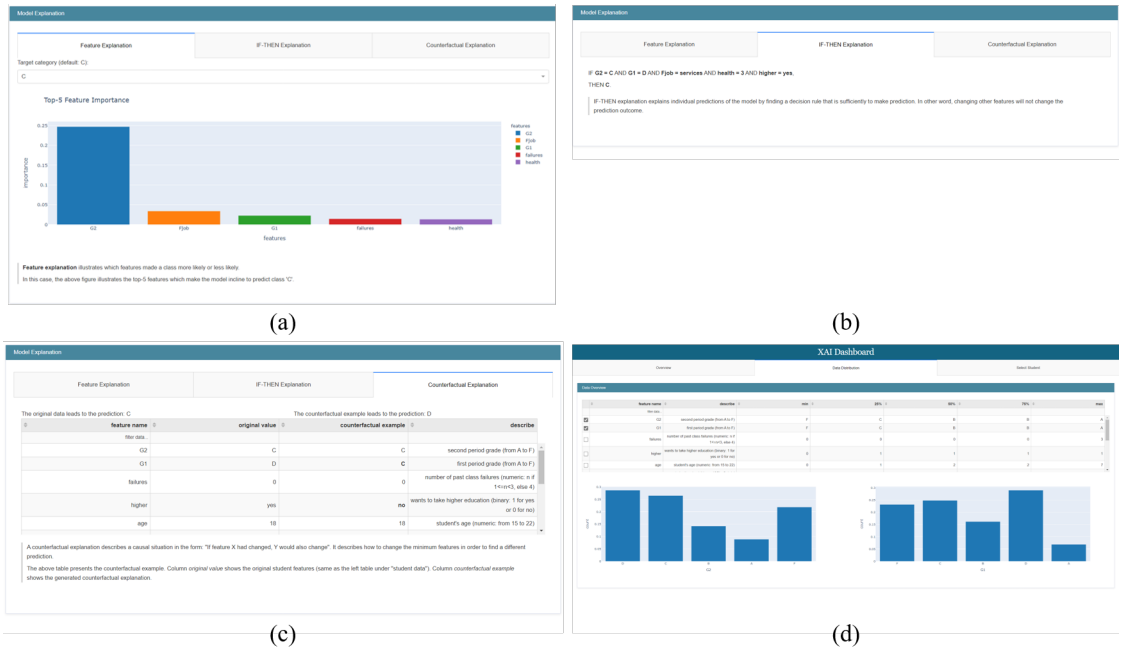

Figure 2: Explanations of model's predictions and data distribution. We provide three different model explanations: (a) the feature-based explanation shows the five most important features for the certain model predictions; (b) the IF-THEN explanation illustrates sufficient conditions to be hold for a certain model prediction; (c) the counterfactual explanation shows counterfactual examples in which lead to different model predictions. We also show histograms for each features in (d) to illustrate the data distributions.

preferences and issues users had when using the model, and **the impact of their prior experience on their understanding** of the ML model.

Table 3: An overview of the interview protocol investigating experience of using the ML interpretability tool.

Is this your first time to use a machine learning model?
What challenges did you have when interpreting the feature-based explanation?
What challenges did you have when interpreting the IF-THEN explanation?
What challenges did you have when interpreting the counterfactual explanation?
Did you understand the three explanation styles? Which parts are or not understandable to you?
Which explanation do you prefer? Why do you prefer this one?
What kind of features do you think are important to predict the students' final scores? Why?
What other features do you want to add to predict students' final scores?
Do you think the model is accurate enough for you to predict the students' final scores?
Did the data distribution help with your decision in predicting the student's final score?
Based on your answers, why do you disagree with the model prediction? (if there are disagreements)

4 Findings

The machine learning interpretable model was designed for users to predict students' final scores based on a series of features such as two exam scores, parents' education and jobs, free time, weekend alcohol assumption, etc. Through analyzing how the two different groups (Teachers vs. ML practitioners) use the model, we understood the common issues and preferences they had, enabling us to generate useful design implications to improve the model or the dashboard to make it more understandable to both teachers and ML practitioners. The Table 2 below shows the demographic information of the four users. *User A* and B belong to Teacher Group, C and D to Machine Learning Group.

4.1 Preferences and Challenges with the Explainable ML Systems

The preferences refers to which explanation users preferred when using the model to predict students' final scores. According to the interview data, we found that, among the three explanations, all of the users preferred the first explanation feature-based explanation whatever their education backgrounds and prior experience are. All of them indicated that this explanation is clear and straightforward to understand (Figure 2, (a)), and they feel comfortable in using it when compared to the other explanations (IF-THEN, Counterfactual).

> *Researcher: "Which explanation do you prefer?"*
>
> *User A: "I think (I prefer) feature explanation because this one makes more sense to me. I think the feature-based explanation is the easiest one for*

me to understand. Because at first I didn't really understand this model, but then I kind of understood this one because the five features are really clearly demonstrated here and I can see the relation between this model and the values students will get."

User B: "I think that it is really helpful. The feature-based. It really helps me to judge the grades."

User C: "The feature-based explanation is very easy to understand, and you can compare different numbers and see which feature is most important among the five top features to predict students' scores."

User D: "That I suppose that the first explanation I feature makes me feel the most comfortable."

However, we found that most of the users had problems with the IF-THEN explanation and the counterfactual explanation. Three users complained that the IF-THEN explanation is complicated and because it includes many features to determine some students' final scores. For example, for the # 10 student Gloria, there are 9 features used to predict her final score (see Figure 3). One user said that she did not understand the logic behind the IF-THEN explanation. In other words, why the IF-THEN explanation picks some specific features instead of others to make a prediction. So she would not choose to trust this explanation to predict students' scores.

Researcher: "What challenges did you have when interpreting the IF-THEN explanation?"

User B: "Yeah, I know the rule for this model (IF-THEN explanation), but It doesn't make sense for me. Since it is like for the student number 10, the conditions are too much. There is a correlational relationship between the conditions with the prediction, but not causal relationship. This explanation cannot convince me."

User C: "For different students and it will show like the if-then conditions will be different and like for example, like now I choose this the 10th students, Gloria. Yeah, and it will show so many conditions under IF-THEN explanation and I need to like, look, look back to the student data one by one. Yeah, and I think because different students have different conditions. So I don't really know the basic logic of this model. So I think it's maybe a bit how to say it a bit more complicated."

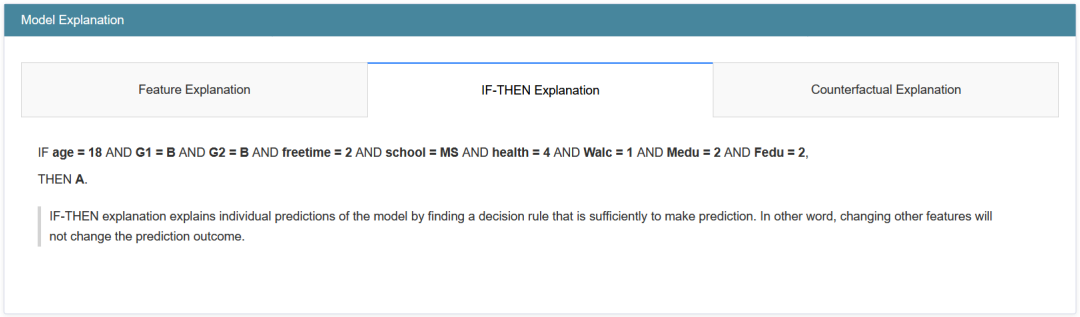

Figure 3: The IF-THEN explanation for the tenth student.

Besides, all of the users felt confused about the counterfactual explanation. They could not understand how the counterfactual explanation works. All of them said they did not use the counterfactual explanation to predict students' final scores because they do not understand this explanation very well.

> *Researcher: What challenges you have when interpreting the counterfactual explanation?*
>
> *User A: "To be honest, I don't quite understand this, this model. (counterfactual) because I don't I don't understand why this value change that value must change because of it."*
>
> *User B: "Oh, and the counterfactual explanation me. Maybe. Maybe. I don't understand that. So I ignored that part."*
>
> *User C: "Like I mentioned before, some of them are not that correct. For the first student, you say the original value for G2 is D and the counterfactual example is C, so when the, when the grade gets better, the predicting the prediction gets lower, so I don't know why this happened."*
>
> *User D: "Um, cause the counterfactual explanation like the original value is like is the same as the left at this table student data and the third, the third column counterfactual example. It's not the full opposite to the original data, it just like a random like that I see it already described a casual situation. Yeah, so like some some some data in the counterfactual example would also be the same as original data and. How to say, I don't, I don't know why but yeah i i don't i don't like this too much like."*

In a nutshell, all of the users preferred the feature-based explanation no matter what their backgrounds are, and they mentioned that this explanation is straightforward and

clear to understand. This indicates that feature-based explanation is more user-friendly for lay people even though they do not understand machine learning.

4.2 The Impact of Users' Prior Experience on Their Understanding of the ML Models

In this study, we asked two different groups of people to use the machine learning model, one with K-12 teaching experience but without a machine learning background, while the other group had machine learning experience but no teaching experience. Through performing the inductive thematic analysis, we found that the users' prior experience or their education background influence their understanding of the model and how they use the model to predict students' final scores.

For example, both users in the Teacher Group thought that some features were more important than others to predict the students' final scores. They mentioned that the two exam scores, students' motivation, absences, and attitude towards learning are important features. Based on prior teaching experience, both teachers agreed that the parents' education and jobs are not important to determine a student's final score, because they thought that it is the students themselves who determine their studies, not their parents, in particular when the students are old enough. Opposite to teachers' opinions, in the ML Group, both machine learning practitioners assumed that the parents' education and jobs are important to be considered when making a prediction on students' final scores. When asked about why they think so, both of the ML practitioners said it was based on their personal experience. It is interesting to find that teachers and ML practitioners held different views on the importance of parents' education and jobs in predicting a student's final score.

> *Teacher Group:*
> *"Just like I mentioned before, the father's job and for those group Like 17 and 18 so dear father, and the mothers may provide some support for students. But they can accept or not. So I don't think these factors from parents are very important."*
>
> *"Based on my prior teaching experience, I had a student whose parents are professors, but the student's performance is very bad. I think it is more dependent on the students' themselves instead of their parents for their studies"*

ML group:
"You know, based on my personal experience and from the internet information, I feel parents' education is important."

In addition, when asked about what other features they wish to add to predict students' final scores, four users conveyed completely different ideas. For example, one user thought that the difficulty of the exam was an important feature to predict a student's score. The easier the exam is, the higher score a student can get, and vice versa. Another user assumed that the mood of the students who take the exams was important. Besides, other features such as teacher-student relationships, whether living on or off-campus were also regarded as important by the users. It is noticeable that each user has different views on the features they wish to add based on their personal experience. Therefore, it is hard to cater to everyone's needs when designing the machine learning model.

To summarize, we found that the Teacher Group held highly consistent views on the *existing features* which they regarded as important or trivial to predict students' scores. For example, both of the teachers believed that students' motivation and attitude towards studies are critical features determining their scores, while parents' jobs and education are not important. This finding is inspiring to us because it reflects the perspectives of teachers in evaluating a student's performance. If we want to introduce the machine learning model to more teachers in the future, we need to take account of their ideas into our ML model design.

4.3 Design Implications

According to the thematic analysis of our interview transcript, we identify several design implications for developing an explainable machine learning system.

First, we find out that all of the invited users prefer feature-based explanations. The feature-based explanation is simple and easy to understand, even for lay people who do not understand machine learning. *User C* also leveraged this explanation to identify two false predictions by the model. We believe that a good explanation should not only breed trust in users but also determine when to trust the model's prediction or not, and the feature-based explanation satisfies both requirements in this system.

Second, it is necessary to constrain the complexity in the IF-THEN explanations. Some of the users would like to try the IF-THEN explanation, but all of them reported that they were more reluctant to use this explanation when the if conditions become uncontrollable. *User B* and *User C* both noted that they would skip the IF-THEN explanation in student

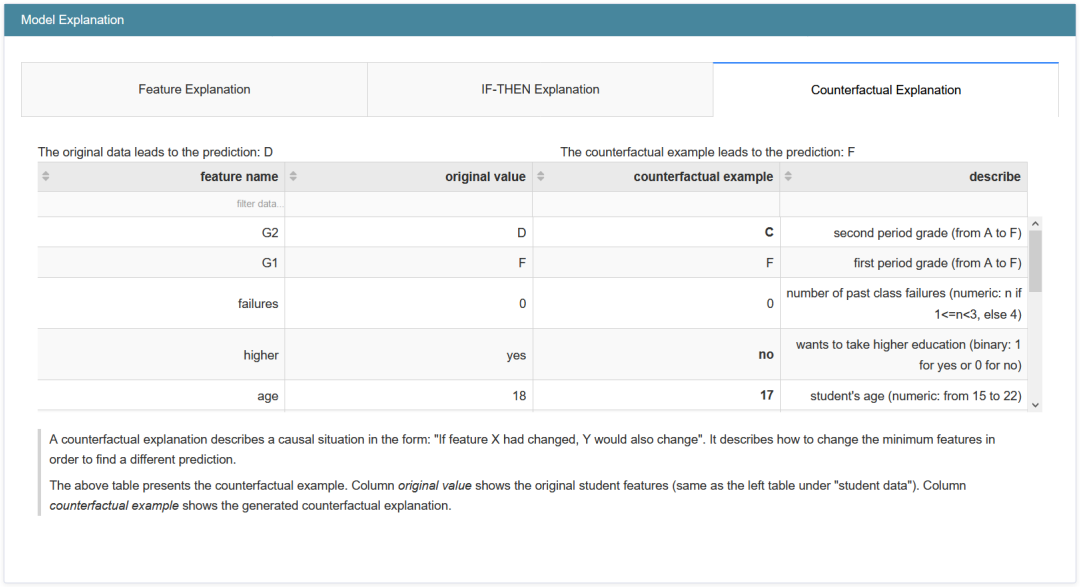

Figure 4: The counterfactual explanation for the first student.

#10 due to too many conditions (nine in total, see Figure 3). The design implication is that the complexity of the explanation will undermine the initiative in using the explanations. This finding is consistent with Narayanan et al. (2018)'s study, where they reported the model complexity increased the response time while had little impact on the accuracy in making the decisions.

Third, none of the users prefer to use counterfactual explanations. This finding is not consistent with previous literature, such as (Binns et al., 2018), where they suggested that counterfactual explanations were preferred for humans in making decisions. Two users in the teaching group cannot understand the counterfactual explanation because "why this value change that value must change because of it". They felt the counterfactual explanation was rather random. Two users in the machine learning group can understand the counterfactual explanation, but they did not trust the explanation due to the adversarial examples. *User C* noted such abnormality in the counterfactual explanation of the first student; the explanation suggests that if the second grade of the test increases from D to C, the student will actually get an F who was predicted to get a D. This phenomenon results from the model's vulnerability, and the counterfactual explanation serves as the adversarial example to identify the model's weakness. We believe that it is crucial to prevent adversarial examples in making the counterfactual explanation. To the best of our knowledge, preventing the adversarial example has not been identified as criteria in the literature in generating the counterfactual explanations.

Fourth, we recognize that cultural context should be taken into account when designing the human-AI systems. Our dataset is collected from two Portuguese schools, while all of our users are grown in China. Such cultural differences lead to certain hardships in understanding the data and explanation. During the user study, three out of four participants asked about the meaning of the feature "freetime". This feature measures the freetime after school. In China, this feature will not be brought up because most students tend to devote all of their time after school to study. However, in most Western countries, some students might also take part-time jobs after school. Therefore, it is necessary to consider the cultural factors when designing explainable systems.

Furthermore, our participants suggest improvements to our online dashboard, which is also valuable in designing an explainable ML system for teachers. One user in the teaching group suggested that showing the trend of the test grade can better judge students' performance throughout the semesters. This user also suggested that it would be optimal to add another test score to see a clearer trend.

Finally, we believe that the personalized explanation is in demand, which might boost productivity in making decisions. Section 4.2 shows that different user groups have distinct views in perceiving machine explanations. Even in the same group, background at the individual level could also impact users' understanding of the model explanations. Therefore, the personalization in explanation can carter to such differences so that the general usability of the system will be improved.

5 Discussion and Conclusion

In this pilot study, we developed an explainable machine learning system and recruited four users to use the system. This explainable ML system is used to predict students' final scores, so we invited two users (Teacher Group) with teaching experience to use this model. In comparison with the Teacher Group, another two users with no background in teaching possess machine learning experience. The aim of this study is to understand the factors influencing the two groups' use of the model and the challenges they had.

Through conducting thematic inductive analysis, we found that all of the users preferred the feature-based explanation more than the other two explanations. Nearly all of them thought that the IF-THEN explanation was complex, and the counterfactual-based explanation was hard to understand. Moreover, we also identified that users' prior experience and their background had impacts on their understanding and use of the model. For example, teachers held similar views on the features which they regard as important or trivial to determine students' final scores.

By doing data analysis, we developed a good understanding of users' experience in using the ML model, helping us generate some design implications aimed at upgrading the current ML model to make it clearer and more understandable to users.

Our work has several limitations. First, the sample size is small because we only invited four people to use the ML model. In the future, we plan to recruit more people via Amazon Mechanical Turk to use the ML model. Second, all of the users are originally from China, and their understanding of the model might be affected by the cultural context. For example, weekend alcohol (one of the features) is a rare phenomenon in China, and most of Chinese teachers will not consider it when predicting students' final scores. Therefore, we need to consider the cultural context of the users when designing a ML model for them. Third, the size of our dataset only contains roughly 600 samples, which is not enough to train a highly predictive machine learning model. Because of the limitations in the sample sizes, the ML model might be susceptible to adversarial attack encountered in the our study (see discussions in Section 4.3).

Overall, we study two groups' use of the ML interpretability tool and conducted pilot interviews (N=4) to uncover the common preference and challenge faced by all of the users and their personalized understanding of the features. In view of the users' experience in the ML model, our study yielded some meaningful design implications which hopefully will help improve the model to make it more understandable to a larger audience in the future.

References

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. (2018). 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Caruana, R., Kangarloo, H., Dionisio, J. D., Sinha, U., and Johnson, D. (1999). Case-based explanation of non-case-based learning methods. In *Proceedings of the AMIA Symposium*, page 212. American Medical Informatics Association.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. In *Advances in neural information processing systems*, pages 8930–8941.

Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Kanehira, A. and Harada, T. (2019). Learning to explain with complemental examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8603–8611.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., and Doshi-Velez, F. (2019). Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 59–67.

Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684.

Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3):31–57.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Mahajan, D., Tan, C., and Sharma, A. (2019). Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*.

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

Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., and Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *AAAI*, volume 18, pages 1527–1535.

Ribera, M. and Lapedriza, A. (2019). Can we do better explanations? a proposal of user-centered explainable ai. In *IUI Workshops*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Van Looveren, A. and Klaise, J. (2019). Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*.

Voigt, P. and Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15.

A Demographic Survey

1. Describe your education background. eg., undergrad major, graduate major.
2. Did you take any education-related courses? For example., teaching methods, or any other education-related courses?
3. Did you have any teaching experience? If yes, how long did you teach before? Which grade did you teach?
4. Did you ever use any data analysis tool? Could you describe how you use the data analysis tool? eg., excel, SPSS
5. Did you have machine learning (ML) experience? What experience do you have (course, project, industry intern, research)? If yes, how long did you use ML before?
6. Did you have a statistical background? If yes, what level of knowledge do you have? What statistics courses did you take? What was the course level?
7. What mathematical background do you have? What mathematical course do you take? What is the course level?

B Individual Contributions

Hangzhi Guo:

- Developed the online dashboard to present the explainable ML system
- Conducted the user study and interviews
- Weekly meet with the peer to discuss about the data analysis and paper writing
- Wrote the report Section 2 (Literature Review), Section 4.3 (Design Implications)
- Reviewed the remaining report

Na Li:

- Designed survey and interview questions
- Conducted interviews with the participants
- Analyzed interview transcripts
- Weekly meetings with peer to discuss about the data analysis and paper writing
- Wrote the paper abstract, introduction, findings, discussion and conclusion. Helped reviewed the other parts peer wrote.