



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

REPORT

HYBRID HUMAN AI RESEARCH STUDY TRANSPARENCY, INTERPRETABILITY, AND ACTIONABILITY OF EXPLAINABLE AI IN EDUCATION

EPFL COURSEWORK
MASTER SEMESTER PROJECT IN
MACHINE LEARNING FOR EDUCATION LABORATORY

Juliette Parchet
(Supervisor: Vinitra Swamy)
(Head of the Laboratory: Tanja Käser)

1st January 2024

ABSTRACT

This research paper explores Explainable Artificial Intelligence (XAI) within education, focusing on MOOCs' settings (Massiv Open Online Courses). In recent years, the integration of AI (Artificial Intelligence) into educational settings has significantly transformed traditional learning paradigms. XAI tends to provide more transparent decision explanations, crucial in the context of predicting student success or failure. This study first investigates through a semi-structured interview XAI model inconsistencies' impact on user trust and decision-making. It also delves into educators' level of explanation needs within MOOCs, among consistency, accuracy, real-time, actionability, and human-interpretability. In a second phase, this study designs and assesses through a survey explanation communications for a black-box predictive model, in the context of MOOCs. In particular, it examines the transparency, human-interpretability, and actionability of these designs, by addressing two key elements in XAI: format and completeness.

CHAPTER 1

INTRODUCTION

1.1 CONTEXT

Explainable Artificial Intelligence (XAI) has become a pivotal tool at the intersection of AI and education, allowing AI systems to offer understandable and transparent explanations for their decisions. This capability enhances human comprehension and trust in these systems, which is crucial as education increasingly incorporates AI-driven tools and platforms. Practical applications of XAI models have recently emerged in Massive Open Online Courses (MOOCs) [Swamy, Radmehr et al. 2022], designed for massive participation and open access via the Internet. MOOCs typically provide features such as video lectures, interactive quizzes, forums, and peer-to-peer learning, making them fertile ground for building models that can predict student success or failure. However, merely predicting outcomes without understanding the reasons behind them offers limited utility to educators. Herein lies the potential benefit of an XAI model, which can not only predict but also explain how the features influence the prediction. This information would empower teachers to take targeted actions and potentially alter a student's academic trajectory.

1.2 MOTIVATION

To ensure the effectiveness of XAI models in education, comprehensive evaluation is essential. Typically, the evaluation of XAI models involves consideration of four criteria, as outlined by Haque et al. [Haque, Islam and Mikalef 2023]: format (the representation format of the explanation), completeness (ensuring the explanation contains all necessary information, including supplementary details), accuracy (providing information on the accuracy of the explanation), and currency (ensuring the explanation includes up-to-date information). However, meeting these criteria does not guarantee practical usability for human educators. Haque et al. [Haque, Islam and Mikalef 2023] highlight five significant effects that XAI models can have on humans: trust (the user's confidence in AI systems), transparency (the extent to which AI systems understandably reveal their decision-making process), understandability (how easily users can grasp and interpret AI-provided explanations), usability (the ease of use and practicality of the explanation system), and fairness (ensuring ethical and unbiased treatment of users). Moreover, human-interpretability has also been shown to be an essential criterion [Swamy, Frej and Käser 2023]. Finally, actionability is crucial [Hostetter et al. 2023] to evaluate how XAI could translate explanations into actionable steps that improve teaching experiences.

While the aforementioned criteria list may seem extensive and meticulous, it is not exhaustive. For instance, XAI models exhibit inconsistent predictions, which could be interesting to investigate regarding their effect on humans. Additionally, the list can be challenging to navigate in terms of prioritization. For

example, the discourse surrounding XAI emphasizes the critical significance of transparency [Rachha and Seyam 2023], which ensures that AI decision-making mechanisms are visible, thereby enhancing trust and collaboration between human users and AI systems.

In light of the contextual background provided regarding XAI in education, the motivation behind this research project sharpens. The research in the field of XAI in education appeared to reveal a shortage of critical attributes : transparency, human-interpretability, and actionability. We aim to contribute to the advancement and refinement of communicating black-box predictive models' explanations tailored explicitly for educational contexts. We seek to create and assess XAI methodologies focusing on transparency, human-interpretability, and actionability of XAI in education. This involves acting on how explanations are presented by working on the format and completeness of the communicated explanations.

1.3 GOALS

This research project starts by assisting the investigation of the influence of inconsistency on XAI users engaging with black-box predictive-models within XAI frameworks in MOOC educational settings. Indeed, apprehending the impact of inconsistency on XAI users is crucial to understanding how varying or even contradicting information within AI systems affects user trust, decision-making, and reliance on XAI.

Following that, we assist in the exploration of the requirements for explanations among educators engaging with black-box predictive-models within XAI frameworks in MOOC educational settings. These requirements encompass accuracy, consistency, real-time, human interpretability, and actionability. It is essential to analyze the explanation needs of educators to effectively tailor XAI systems within educational contexts. This ensures that these systems meet the specific requirements and challenges of the educational domain. Understanding educators' needs helps in creating explanations that align with their teaching methodologies. It also tends to improve the adoption and integration of AI technologies in educational contexts.

Finally, we focus on enhancing transparency, human-interpretability, and actionability of black-box predictive model explanations by addressing two dimensions of XAI: format and completeness. The goal is to design and assess different ways to communicate these explanations, such that learning experts would find these designs actionable, transparent, and easily interpretable.

CHAPTER 2

METHODS

In the initial phase of this research project, semi-structured interviews are conducted to study key questions regarding XAI in educational contexts (the complete interview can be found at [Vinitra Swamy 2023]). In the first part of the study, the participants are told that they are an instructor of a large course with more than 100 students, so unfortunately they are unable to interact with all of their students individually. As an instructor, they are interested in identifying the “struggling students” in their course to provide additional support and materials. The study presents two teaching scenarios to participants: one involving two different human teaching assistants (TAs) and another with two computational models with the same architecture, same model accuracy but which have different predictions and/or insights. Participants are asked to assess believability, trust, and intervention indicators in these scenarios. Each participant considers scenarios involving three students in both TA and model contexts, observing escalating levels of inconsistency between the two TAs/models with each student. The interviewees must evaluate trust and intervention indicators based on explanation quality and consistency. At the end of the first part of the study, participants are asked to reflect on changes in trust levels between TAs and models, to get deeper insights into the difference in perceptions and trust between the scenarios.

In the study’s second segment, participants engage with three distinct scenarios. Initially, they assume the role of a school principal tasked with making annual policy decisions for the entire school using a predictive model. Next, they embody a tutor meeting a single student weekly, aiming to personalize learning based on the predictive model’s insights into the student’s weaknesses. Lastly, participants take on the persona of a middle school teacher conducting frequent meetings with a group of students, using a predictive model to craft interventions tailored to a group of students displaying similar behavior.

These interview answers were diligently transcribed, focusing on questions such as “Which XAI methodologies do educators trust?”, “To what extent are learning science experts affected by the inconsistency of explanations offered by XAI systems?”, and “What needs do different educational stakeholders have?”—covering a scope of five essential needs: human-interpretability, real-time responsiveness, actionability, consistency, and accuracy. Following the interview transcriptions, a comprehensive analysis was done to examine and interpret the interviewees’ responses (to see full code: [Parchet 2023]). Initially, a demographic analysis was conducted to gain a comprehensive understanding of the diverse educational backgrounds of the learning science expert interviewees. Subsequently, participants’ responses, encompassing preferences like trusted experts and chosen intervention features, were analyzed using basic frequency counts and ratios and then translated into visual representations. Thirdly, to analyze the interviewees’ answers to the question on changes in trust between TAs and models, a coding scheme was formulated (see Table 2.1), based on the coding scheme of Nazaretsky et al. 2022

Category	Short Description	Detailed Description
Accuracy & Reliability	1 - TAs have human characteristics	TAs use human, personal connections with students: 1) TAs understand social, emotional, and motivational factors important in teaching. 2) TAs know students' history outside the system. 3) Models are more objective vs. TAs are more subjective.
	2 - TA experience	TAs have more experience: 1) TAs have real-life experience and human intuition. 2) AI developers lack pedagogical and educational experience.
	3 - Trust in TA from personal interactions	Know TA on a personal level, informed of their strengths and weaknesses.
	4 - Models are objective, TAs are not	TAs are more likely to make mistakes than models — models are objective.
	5 - Models do not have a complete picture	Features passed to models do not cover a full holistic picture.
Technical	6 - Models can scale	Models can help more students than TAs: 1) Models are more efficient (less time per student). 2) Models are more effective (no errors due to tiredness). 3) Models are scalable (more students helped).
	7 - Models cannot give more detailed info	Models cannot give more information about a situation beyond what their training data allows, but TAs can.
Affective	8 - TA disagreement is acceptable	Disagreement between TAs is considered a regular, valid situation. When models disagree - something is wrong with the models.
	9 - TAs have empathy and compassion toward students	TAs can decide how not to negatively impact or overwhelm students.
	10 - Models are more detailed than TAs	Models have more detailed information than TAs. This enables models to have more granular insights than TAs.

TABLE 2.1

Coding scheme for perception of TA vs. model responses, based on the coding scheme of Nazaretsky et al. 2022.

In the second part of this research project, we focus on enhancing transparency, human-interpretability, and actionability of black-box predictive model explanations in a MOOC setting. We do that by addressing two dimensions of XAI: format and completeness. The goal is to design and assess different explanation communications such that learning experts would find them actionable, transparent, and easily interpretable. In this order of ideas, seven different ways to communicate explanations are conceived—three visual and four textual (find the code at [Parchet 2023]). To reduce the massive amount of data given by the XAI model, only the top five—over 82—impactful features for the student (toward failure or success) are kept. This streamlining reduces completeness in favor of improving human interpretability, enabling quicker comprehension, and encouraging actionability. Each feature includes a weekly value, importance score, and class average. Since we focus on a mid-semester intervention, we limit the data to the initial five weeks out of ten. Nonetheless, we could easily extend this intervention to earlier or later stages in the semester. Below are the design specifics of the seven explanation communications.

Visualization 1, shown by Figure 2.1 helps us understand the difference between the student's features score and the class's average performance through a bar chart. Bar chart format was chosen for the first visualization, which we tried to keep as simple and understandable as possible. [Meloncon and Warner 2017] regards bar charts as appealing and easily comprehensible in terms of cognitive simplicity. However, while bar charts are attractive and straightforward, they convey limited information and lack the timeline intuitiveness of line plots.

Visualization 2, shown by Figure 2.2 helps us understand which features contribute to the student's predicted success and failure, and how they are performing compared to the class average. For this visualization, we chose the line plot design that highlighted effectively the timeline of the features. Compared to Visualization 1, we added complexity by introducing the mean score of the class for each chosen feature.

Visualization 3, shown by Figure 2.3 helps us understand which features contribute to the student's predicted success and failure, and how they are performing compared to the class average. Visualization 3 enhances Visualization 2 by incorporating the importance score dimension as the dot size. [Swamy, Radmehr et al. 2022] introduces the importance score as an indicator of a feature's significance in

predicting success or failure for a particular prediction. This score spans from a high positive value, signifying substantial importance for success, to a high negative value, denoting consequential importance for failure. It encompasses values near zero, indicating low feature importance.

Three of the four text explanations were automatically generated by chatGPT 3.5 [OpenAI 2023]. These generated text explanations are the result of carefully crafted prompts designed to generate a natural-sounding data scientist report, easily understandable by non-experts. This report is based on the same data used for the visualizations: the top five impactful features for the student (value, importance, and class average), over the initial five weeks of the course. Please find all the details of the prompts and generations in the Appendix chapter.

Text Explanation 1's prompt instructs to provide a detailed, feature-specific, and weekly description of the data behavior. This yields a comprehensive 540-word description, well-structured and organized.

Text Explanation 2's prompt instructs to provide a synthesized weekly description of the data behavior. It results in a much shorter 240-word description, not structured by features but only by weeks.

Text Explanation 3's prompt directs to offer a condensed feature-specific description of the data behavior. This renders a 221-word description, organized solely by features, not weeks.

Text Explanation 4 is authored by a learning science expert who looked through Visualizations 1, 2, and 3 and drew some conclusions.

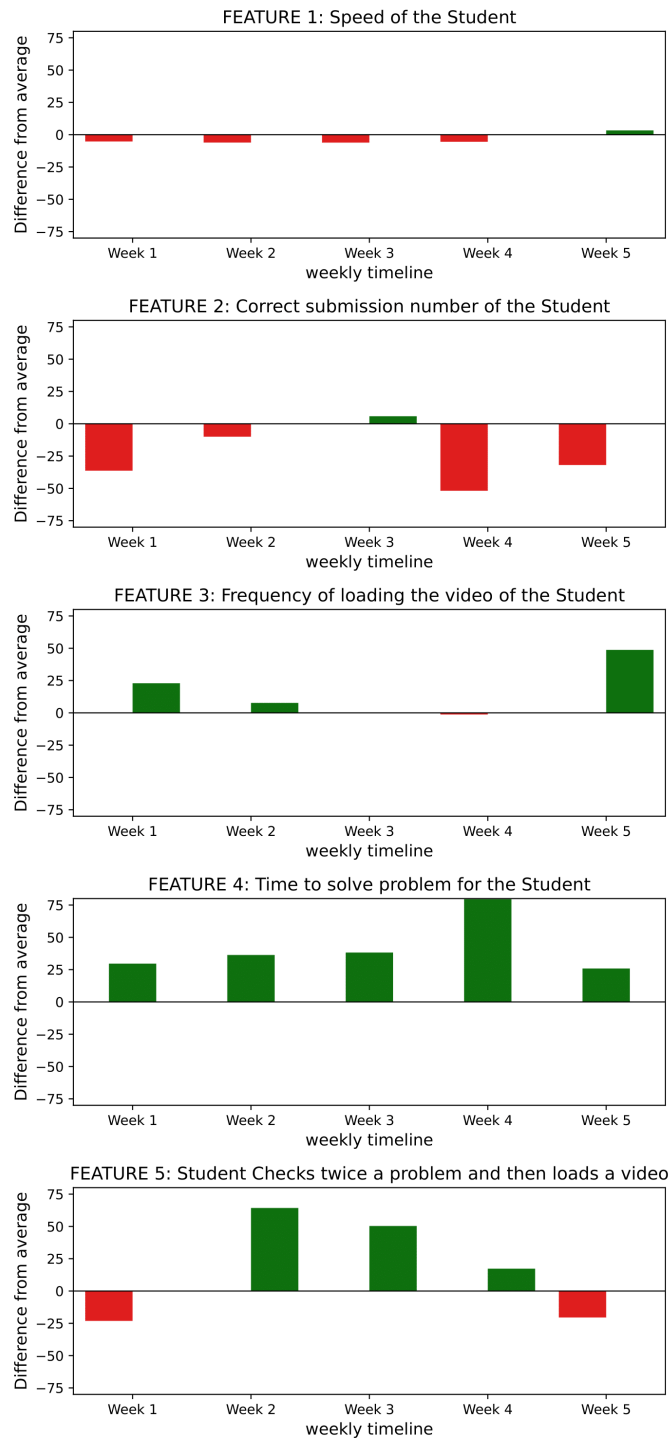


FIGURE 2.1

Visualization 1 communicates the explanations for a failing student prediction halfway through the course.
Red bars mean the feature's value is below average and green bars mean it is above average
The y-axis value represents the difference of the feature value from the mean
The x-axis value represents the student's weekly progression

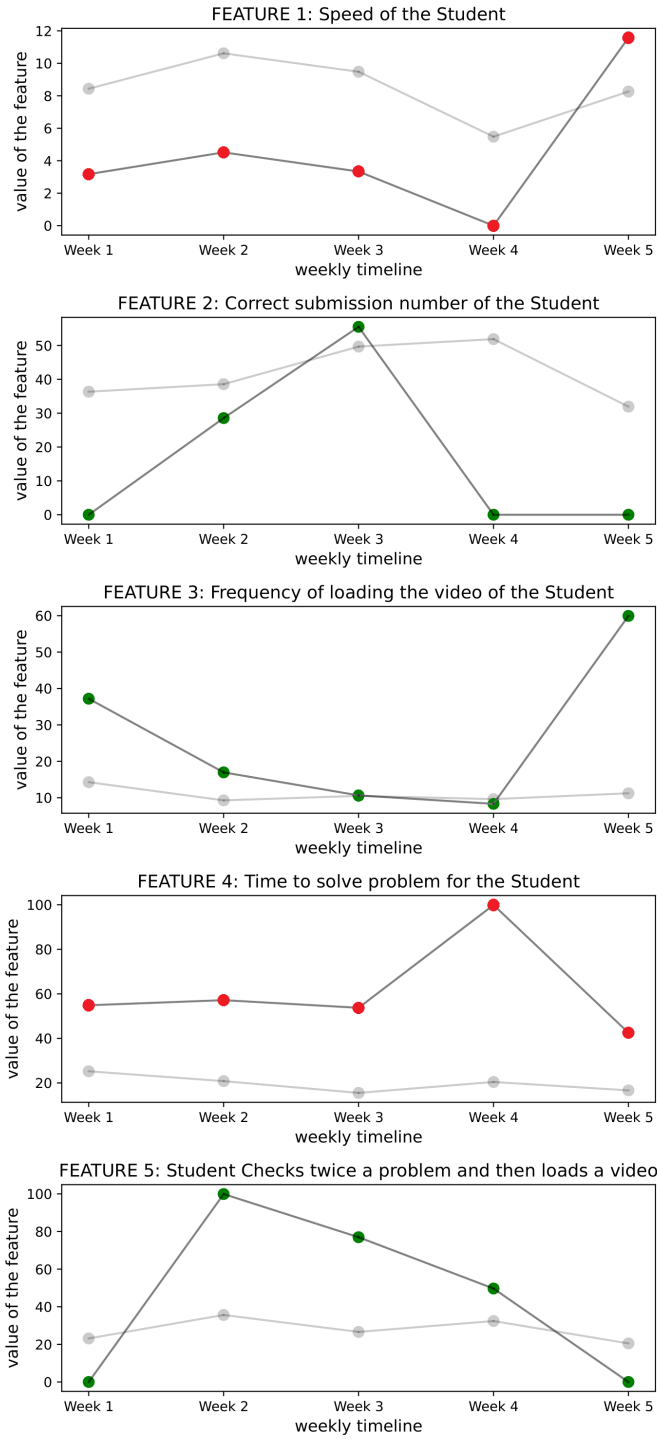


FIGURE 2.2

Visualization 2 communicates the explanations for a failing student prediction halfway through the course.

Red dots represent failing behaviors and green dots represent passing behaviors

The y-axis value (or height of the dot) represents the value of that feature for the student

The x-axis value represents the student's weekly progression

The light gray line represents the average student performance in the class

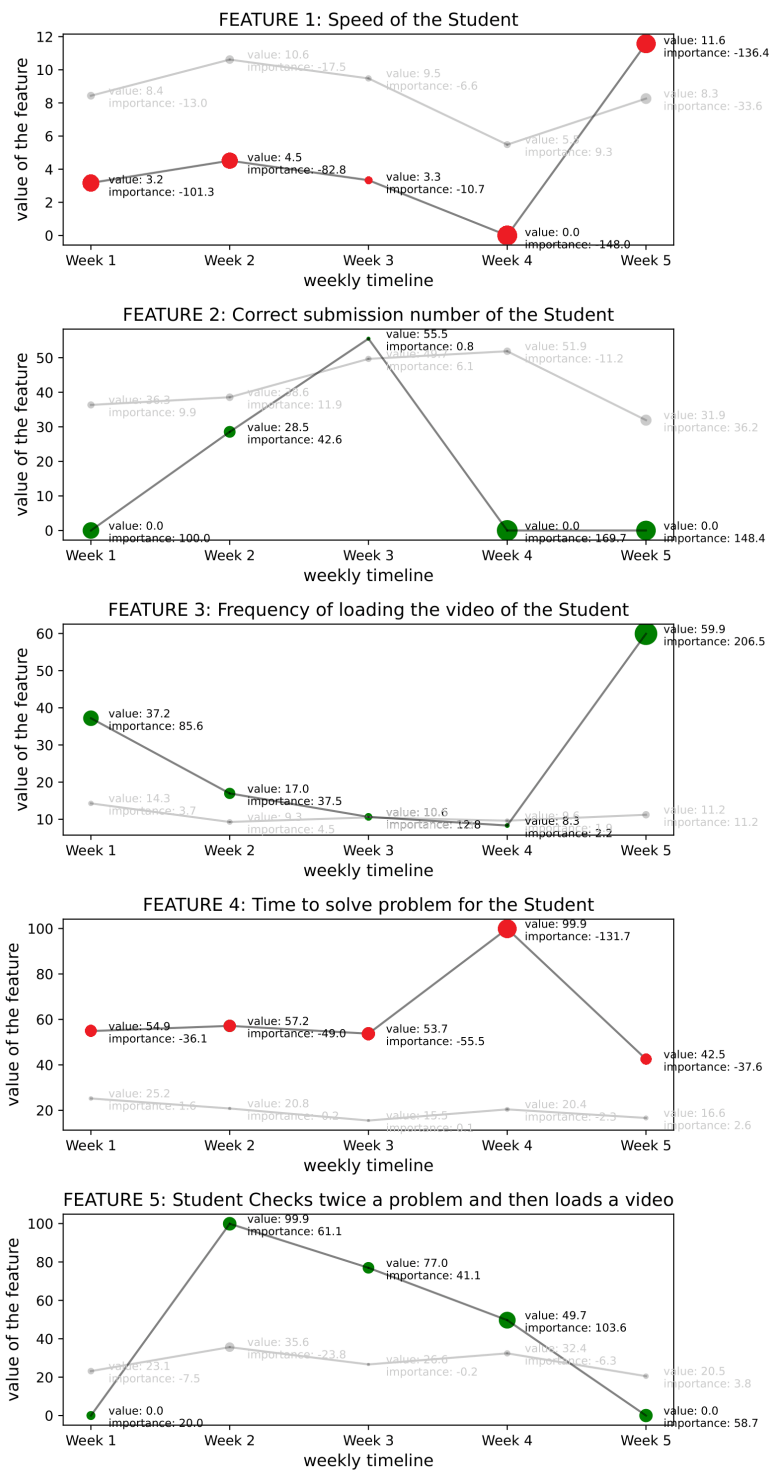


FIGURE 2.3

Visualization 3 communicates the explanations for a failing student prediction halfway through the course.

Red dots represent failing behaviors, green dots represent passing behaviors

The size of the dot, also called the "importance score" represents how important this feature is towards the passing or failing direction

The y-axis value (or height of the dot) represents the value of that feature for the student

The x-axis value represents the student's weekly progression

The light gray line represents the average student performance in the class

In order to assess these different explanation communications, a survey is conceived (find the survey as well as the raw results here [Juliette Parchet 2023]). The questionnaire is around 20 minutes long, meant to be self-explanatory. The study takes place in the context of a black-box predictive model in a MOOC setting. The model predicts at mid-semester if a student is likely to fail or succeed in the course, and gives a predictive grade. Alongside these predictions, the model provides explanations on why the model predicts success or failure. Unfortunately, these explanations are extensive and appear to lack coherence or meaning at first glance. To unravel the complex explanations given by the XAI model, the seven explanation communications designed are presented to the participants. The survey scenario places participants halfway through the semester, where a student is predicted to fail with a grade of 3.5 out of 6. Participants act as TAs or teachers and are tasked with understanding why the student is failing and proposing an intervention, based on the explanation communications. This survey aims to investigate the effectiveness of different explanation formats in aiding teaching assistants during a mid-semester intervention for a struggling student, focusing on actionability, human-interpretability, and transparency.

Let's now review the study's specifics. All explanation communications elicit responses to three questions:

- "What do you think of the complexity of this explanation?"
The answers are graded on a spectrum from "Too simple" to "Too complex".
This question seeks to gauge participants' perception of the human-interpretability of the communicated explanations.
- "Does this explanation cover all the information you need to design an intervention for this student?"
The answers are graded on a spectrum from "Missing crucial information" to "Includes unnecessary information"
This question targets participants' perception regarding the transparency and completeness of the communicated explanations.
- "Would you like to have this explanation when you are making decisions about your students?"
One answer between: "It would make my decision process harder to have it", "I don't need it", "I'm neutral about it", "I would like to have it", and "I really need to have it to make the best decisions for my struggling students"
This question aims to grasp participants' perception of the actionability of the communicated explanations.

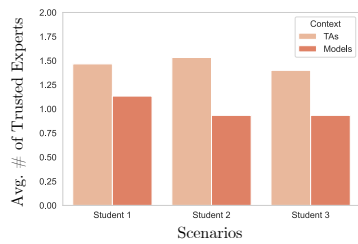
The survey concludes by questioning participants about their preferred explanation format—text, visual, or hybrid . Additionally, participants are asked to provide demographic information.

CHAPTER 3

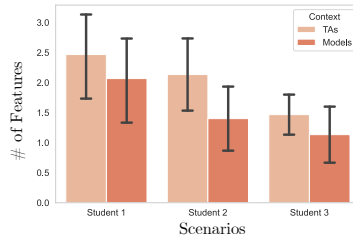
RESULTS

To begin, the findings of the first part of the study are presented (find the code and analysis here [Parchet 2023]). Fifteen participants answered the survey, from various cultural backgrounds and with diverse learning experiences.

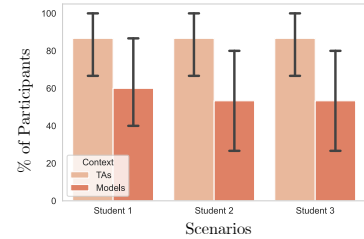
First, we can take a look at Figure 3.1, which displays indicators of trustworthiness, distribution of feature selection, and choice of experts.



(A) Average number of experts found trustworthy across the context of TAs or models for each scenario.



(B) Distribution of the number of features selected for intervention for each scenario.



(C) Percentage of participants who trusted at least one expert for each scenario.

FIGURE 3.1

Trustworthiness (a), number of features selected (b), and choice of expert (c) as answered by the participants for the three different scenarios (small changes in explanation, large changes in explanation, large change in explanation and prediction).

Subsequently, we can take a look at Figure 3.2, which exhibits the attribute scores given by the participants per scenario, for the five attributes: consistent, real-time, accurate, actionable, and human-interpretable.

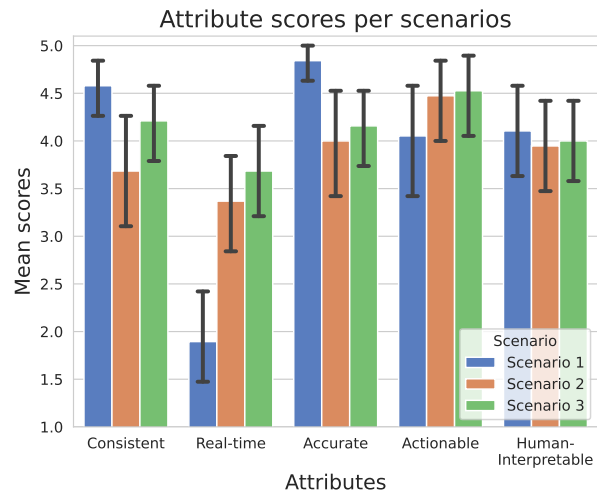


FIGURE 3.2

Attribute scores per scenario, for the five attributes: consistent, real-time, accurate, actionable, and human-interpretable.

Next, the findings of the second part of the study are presented (find the code and analysis here [Parchet 2023]). Ten participants answered the survey, all teaching or assisting in teaching at the Swiss Federal Institute of Technology in Lausanne (EPFL). The years of learning science experience spanned from one year up to ten years.

First, we can observe in Figure 3.3 the responses of participants to the query "What do you think of the complexity of this explanation?".

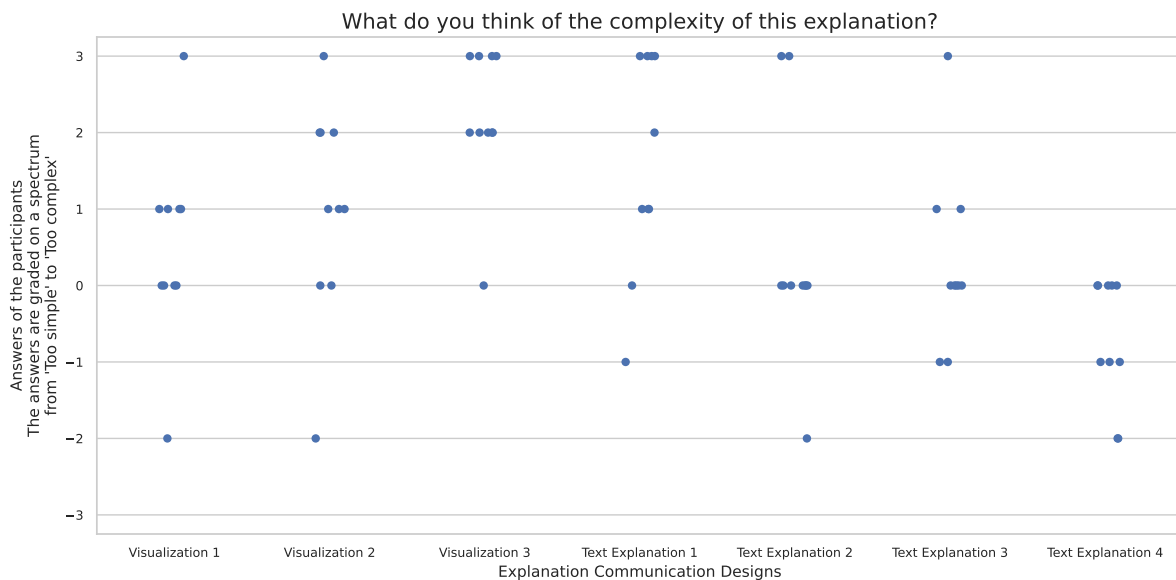
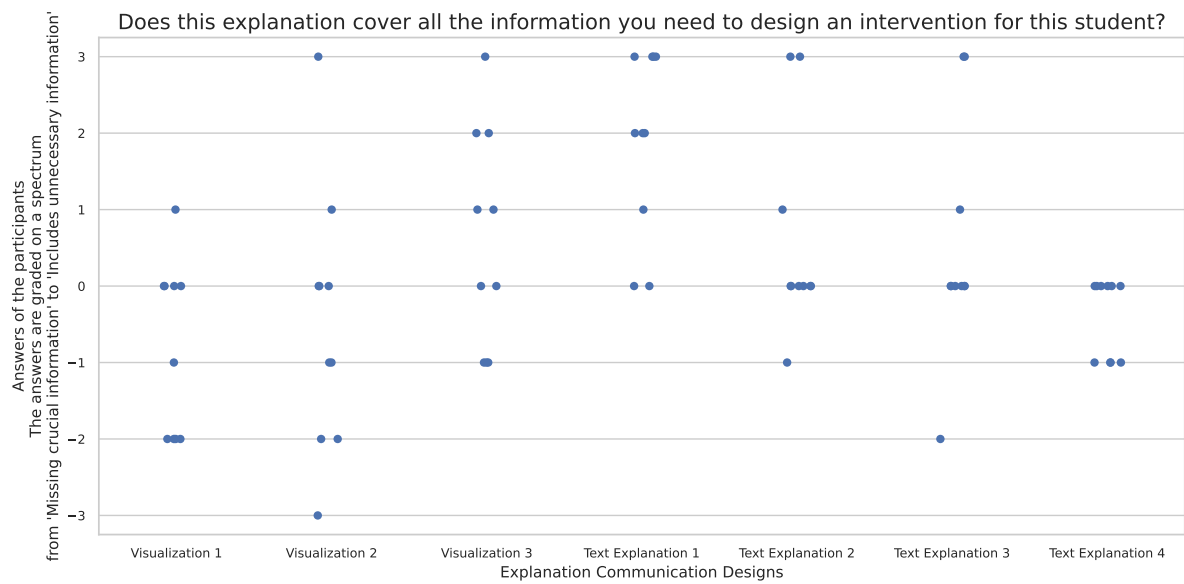


FIGURE 3.3

The participants' answers distribution for the question "What do you think of the complexity of this explanation?". The answers are graded on a spectrum from 'Too simple' (-3) to 'Too complex' (3).

Following that, Figure 3.4 illustrates the outcomes of the question "Does this explanation cover all the information you need to design an intervention for this student?".

**FIGURE 3.4**

The participants' answers distribution for the question "Does this explanation cover all the information you need to design an intervention for this student?". The answers are graded on a spectrum from 'Missing crucial information' (-3) to 'Includes unnecessary information' (3).

Afterwards, Figure 3.5 presents the participants' results of the inquiry "Would you like to have this explanation when you are making decisions about your students?".

Lastly, Figure 3.6 showcases the answers to the question "What type of explanations do you prefer most?". Answers are divided into three categories: Text, Visual, or Hybrid (Both Text and Visual).

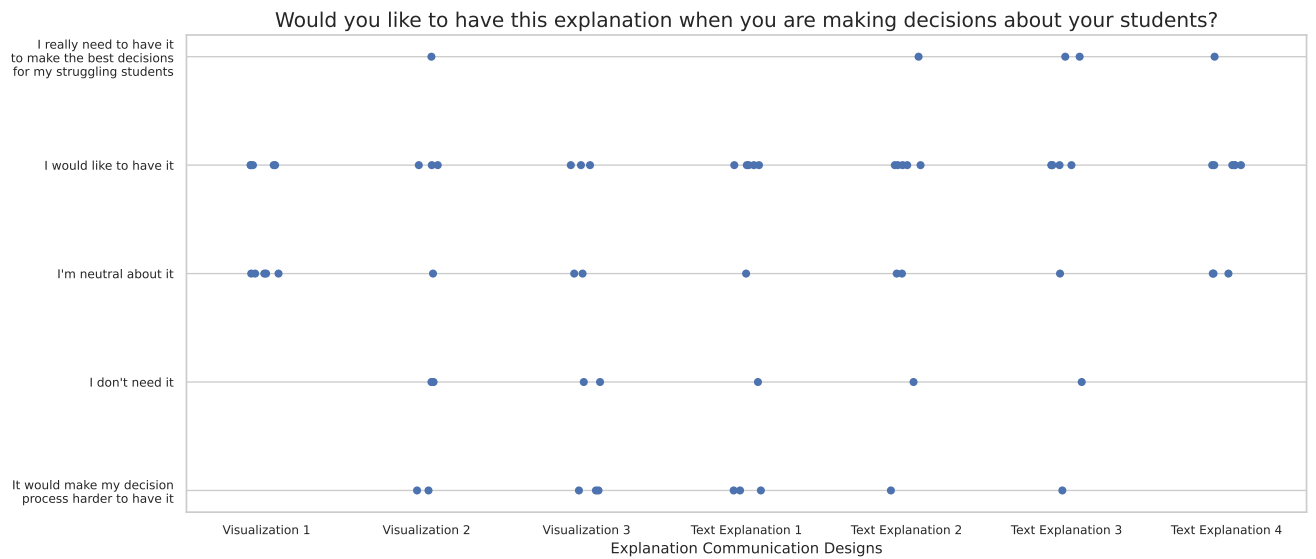


FIGURE 3.5

The participants' answers distribution for the question "Would you like to have this explanation when you are making decisions about your students?". One answer between: 'It would make my decision process harder to have it', 'I don't need it', 'I'm neutral about it', 'I would like to have it', and 'I really need to have it to make the best decisions for my struggling students'.

Which type of explanation do you prefer most?

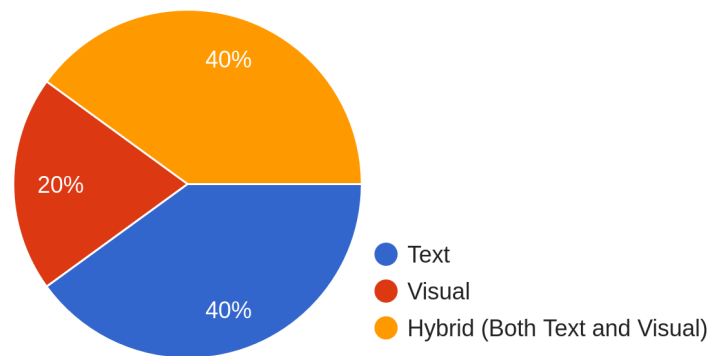


FIGURE 3.6

The participants' answers distribution for the question "What type of explanations do you prefer most?"

CHAPTER 4

DISCUSSION

The discussion of the results starts by diving into the findings of the first part of the research project.

To understand the human perception of and reaction to inconsistency in explanations, we evaluated the answers of the 15 semi-structured interviews with learning science experts. These interviews explored trust in Teaching Assistants (TAs) and models across three scenarios, each demonstrating increasing levels of inconsistency. The subfigure 3.1a shows the average number of TAs (models) trusted by the participants for each scenario. Notably, trust in models decreased significantly with increased inconsistency, contrasting with constant trust levels in TAs across scenarios. The difference implies a greater acceptance of disagreement in human TAs than in models, even with identical information presented to both sources.

Next, we investigate the differences between features selected for intervention between the TA and model contexts across scenarios (illustrated in the subfigure 3.1b). Participants leaned towards selecting more features from TAs, indicating a higher level of trust in human TAs' logic. As explanations and predictions became less consistent, the number of features used for intervention decreased.

The subfigure 3.1c shows the percentage of participants that made an intervention based on features of the information sources they trusted. Participants' intervention choices predominantly aligned with features from the trusted information source, indicating a significant reliance on the expert they trusted. However, there were instances in the model context where participants preferred the model's reasoning for intervention, even without explicit trust in the expert.

In a final analysis, we evaluated participants' answers to the last question of the interview based on our coding scheme (see Table 2.1).

For the Accuracy and Reliability category, 70% of the participants noted the human characteristics and connections of the TAs that are lacking in the models. 20% of participants mentioned the real-life teaching experience of TAs that models do not have, and over 33% of them mentioned that they would trust the TAs more based on their personal interactions with them. One participant said, "If one had more affinity with the TAs and knew how they worked, the trust level might change to higher if they were a good worker or lower if poor worker." Furthermore, 36.67% of participants stated that models were objective and TAs subjective, but for those who mentioned this, they were evenly split on whether model objectivity is a positive aspect. A participant discussing model objectivity stated "With the TAs one cannot expect much objectivity, and one will give, expect, and tolerate more error margin to humans than models. If there are some guarantees that the model is well-trained and consistent, one would trust the model more (but in return, expect models to be coherent in their predictions)." Finally, 40% of respondents (six participants) stated that models did not have a complete picture of student performance, although it was included in the study design that both TAs and models were exposed to the same information and interactions. On the other hand, 6% of participants stated that models were more detailed than TAs (affective factor).

In the Technical category, 10% of participants mentioned the scalability of models as a positive towards their effectiveness and trustworthiness. 16.67% of participants stated the inability of models to give more detailed information about the student situation as a negative, in contrast to the way a TA might be able to answer follow-up questions.

In terms of the Affective category, several participants (13%) stated that TA disagreement is more acceptable than model disagreement, confirming the results from Figure 8a (TODO). 30% of participants mentioned that TAs had empathy and compassion towards students, as a positive attribute supporting human experts. One participant said the following about their (dis)trust of models: “For models, their training, development, limitations, and strong points are unknown. I would trust models more if there was more feedback from learning experts (teachers) on the model in action, and would trust more this kind of validation rather than a classic model validation.”

Lastly, we shift the focus to Figure 3.2, a significant visual representation within our discussion. This figure provides a comprehensive insight into the five fundamental XAI needs, highlighting their nearly equal significance. The slight prioritization among these needs is noteworthy, for it sets aside real-time as less critical than the other needs. The bar chart emphasizes the intricate balance and vital role each of these needs plays within the setting of XAI in education.

We will now continue the discussion of the results by diving into the findings of the second part of the research project.

Upon Figure 3.3 examination, it becomes quite clear that both Visualization 3 and Text Explanation 1 are found rather complex to understand by most participants. Moreover, an analysis of Figure 3.4 shows that most respondents perceived these two designs as containing unnecessary details. These observations strongly indicate that these specific designs suffer from poor human-interpretability, mainly attributed to their strong complexity. An excessive amount of information increases transparency but goes past the point where it's useful. This leads to reduced clarity and makes it harder for the intended audience to use the information effectively. Reviewing Figure 3.5 reveals that the poverty of interpretability and surplus of transparency results in reduced actionability. Few participants considered these designs valuable for decision-making, and a noteworthy portion even perceived these designs as obstructing their ability to take action.

Let's look back into Figure 3.3, this time focusing on Visualization 1, Text Explanations 2, 3, and 4 in our analysis. The majority of participants viewed these designs as suitably complex, indicating good human-understandability. However, referring to Figure 3.4, we note that a significant portion of participants identified Visualization 1 as lacking information. The perceived high human-understandability might stem from a lack of completeness of the explanation communication design, potentially diminishing transparency. Figure 3.5 displays participants' neutral-to-positive inclination towards using this explanation in decision-making regarding their students, signifying moderate actionability. Regarding Text Explanations 2, 3, and 4, Figure 3.4 indicates participants' satisfaction with the balance of information provided. Additionally, Figure 3.5 reflects a substantial preference among participants for these textual explanations in their decision-making processes concerning students. These findings suggest that these textual explanation formats exhibit qualities of being easily interpreted by humans, sufficiently transparent, and positively actionable.

Now let's turn to Figure 3.6. The mixed preferences among participants regarding communication formats in XAI can, in part, be linked to distinct cognitive styles and individual learning preferences. Nevertheless, the significant preference towards textual formats compared to visual ones appears to align with the previous discussion: finding a balance between complexity, transparency, and human-interpretability in visualization is challenging, ultimately diminishing the standalone effectiveness of these designs in facilitating actionable insights.

Figure 3.6 unveils a final significant insight. The 40% preference for hybrid explanations, encompassing both text and visuals, indicates that a significant portion of participants benefit from a combination of textual and visual information. This preference might stem from the advantages offered by each format: the clarity and detail of textual information complemented by the visual aids that enhance comprehension and reinforce key concepts. This finding highlights the importance of providing diverse communication formats in XAI. By offering hybrid modes of explanation, XAI systems can cater to the diverse learning styles and information processing abilities of various users, ensuring better transparency, interpretability, and actionability for its audience.

CHAPTER 5

CONCLUSION

The discussions from the first part of the research project yielded significant insights into how learning science experts perceive inconsistencies in explanations, trust levels in different information sources, and their intervention preferences. The interview showed a human tendency to rely more on the logic of TAs for intervention strategies, preferring their reasoning over the models, especially when faced with inconsistent explanations. Additionally, participants highlighted the human attributes lacking in AI models, such as connections, teaching experience, and personal interactions, that contribute to their trust in TAs over models. The study also revealed concerns about models not possessing a comprehensive understanding of student performance and the need for more validation and feedback from learning experts to trust these models effectively.

In summary, the findings of the first part of the study demonstrate that model and explanation consistency is a large concern for educators. While study participants (learning science experts) tolerated inconsistencies in explanations and predictions from TAs, they were much more sensitive to inconsistencies in model predictions and explanations. Yet, consistency isn't the only crucial need of XAI. Educators view accuracy, actionability, human-interpretability and, to a slightly lesser extent, real-time capability, as equally significant needs.

This brings us to the second phase of the study, where we delve further into the requirements for actionability, human-interpretability, and transparency of XAI. The exploration of diverse explanation communication formats in the second part of the research unveils an interesting trend. Participants demonstrate a clear inclination for textual explanations, attributing their value to high human-interpretability, reasonable transparency, and positive actionability. Conversely, visual explanations face challenges in balancing complexity, transparency, and human-interpretability, affecting their effectiveness in enabling actionable insights.

The want for hybrid explanations, combining text and visuals, suggested a possible approach to address various learning preferences and information processing abilities. This hybrid format seemed to leverage the clarity and detail of textual information, complemented by visual aids that enhance comprehension and reinforce key concepts. It could be an interesting path to look into for complementary research work. Exploring this avenue for further research could offer an intriguing and extensive trajectory that might give improved results on transparency, human-interpretability, and actionability of communicated explanations in educational settings.

CHAPTER 6

REFLEXIVE APPROACH

Analysis and discussion between the book "Weapons of Math Destruction" by Cathy O'Neil and my research work on Explainable Artificial Intelligence (XAI) in education.

6.1 INTRODUCTION AND CONTEXT

With the rise of Artificial Intelligence (AI) over the past decade, the transparency and ethics of algorithms have become significant issues, especially in a fundamental sector like education. That is why, in the context of our research about XAI (Explainable Artificial Intelligence) in educational MOOCs (Massive Open Online Courses), it is relevant to examine the book "Weapons of Math Destruction" (WMDs) by the writer Cathy O'Neil.

In her book, O'Neil's delves into the significant impact of opaque algorithms, or weapons of math destruction, on various aspects of society, including education. O'Neil criticizes the dehumanizing effects of these algorithms, which often reduce individuals to mere data points, disregarding the complex nuances of human experiences. Moreover, she highlights the lack of human interpretability and involvement in model predictions, emphasizing how these algorithms perpetuate pre-existing inequalities and worsen social injustices.

Our study on XAI in education aims to enhance transparency and human interpretability in neural network predictions in MOOC settings. At first, the study investigates the impact of inconsistency in XAI models on user trust and decision-making, acknowledging the importance of understanding how varying information within AI systems affects user perceptions and reliance on AI. Then, the study explores educators' needs for explanations in using black-box predictive models. Finally, the study aims to design and assess different ways to communicate black-box prediction-making neural network explanations such that learning experts would find these designs actionable, transparent, and easily interpretable.

While O'Neil's book largely critiques opaque algorithms across sectors, our research study offers insights into addressing transparency, human interpretability, and actionability challenges within educational AI systems without compromising completeness. Both works highlight the importance of human-centered approaches, advocating for transparency, accountability, and fairness.

6.2 PREDICTIVE ALGORITHMS IN EDUCATION : THE NEED FOR HUMAN-CENTEREDNESS IN AI

WMDs criticizes the deployment of algorithms that prioritize efficiency and profitability over fairness and transparency. The author argues that these WMDs can dehumanize people by reducing them to data points, often making existing inequalities worse. One of the author's points in education was the critique of value-added models used to evaluate teachers, which can sometimes produce inaccurate and damaging outcomes. These models only look at numbers and fail to account for the complicated and varied aspects of teaching and learning. According to O'Neil, one of their default points resides in the lack of human-centric designs in AI.

In our research study, one of our main tasks is to design and evaluate human-centric explanation communications for the sole purpose of making the prediction of a black-box model more understandable, interpretable, and actionable for educational stakeholders. This aligns with O'Neil's concerns about the need for human-centeredness in AI.

6.3 TRUST AND TRANSPARENCY IN AI

A central theme in both works is the necessity of trust in algorithmic systems. O'Neil claims that the lack of transparency in WMDs reduces public trust, as people affected by these models often cannot understand or contest the decisions made about them. This issue is especially relevant in education, where teachers and students may suffer from the inconsistencies of black-box models.

The first part of our research tries to address this challenge by analyzing how inconsistencies in XAI models impact user trust in MOOCs. Through semi-structured interviews and surveys, the study investigates educators' needs for consistent, accurate, real-time, actionable, and human-interpretable explanations. Our research enhances transparency and boosts user faith by designing and assessing explanation communications for a black-box predictive model. By focusing on explainability, our research aims to mitigate the opaque nature of many WMDs described by O'Neil, aligning with the author's concerns about the need for transparency and trust in

6.4 IMPLICATIONS FOR THE FUTURE OF EDUCATION STEMMING FROM WMDS AND OUR RESEARCH STUDY

Both works offer insights into the future of education. O'Neil's critique of WMDs highlights the risks associated with uncontrolled algorithms, encouraging stakeholders to prioritize transparency in educational contexts. Her work acts as a warning, prompting educators, policymakers, and developers to consider the ethical dimensions of their data-driven initiatives.

In contrast, our XAI research tries to pave the way for more human interpretable, transparent, and actionable explanations emerging from black-box predictive models. It presents a framework for leveraging technology to enhance, rather than compromise, the educational experience, emphasizing the importance of ethical and human considerations when deploying AI tools.

6.5 CONCLUSION

In conclusion, Cathy O'Neil's "Weapons of Math Destruction" and our research on XAI in MOOCs both address the essential issue of algorithmic transparency in education. While O'Neil highlights the potential harms of opaque and profit-driven algorithms, our XAI research offers a constructive approach to building

trust through explainability. Together, these works emphasize the need for transparent and human-centric design in AI systems, particularly in the context of education, where the stakes are high, and it profoundly impacts people's lives. As engineers, it is clear that our role in developing AI technologies must include a strong duty to create models that are transparent, interpretable, and fair to users.

CHAPTER 7

APPENDIX

7.1 TEXT EXPLANATION 1 GENERATION

Here are the interactions with ChatGPT 3.5 to generate the Text Explanation 1 for student id 206.

User :

You are a talented data scientist and learning scientist. Bellow are some features, with their value, importance scores, and class average. You want to take the following data to generate a natural sounding, short, concise explanation of the ungendered student Charlie performance for a mid-semester intervention. Positive scores mean that the feature corresponded to student success, negative scores means that the feature contributed towards student failure. High value means the feature was observed a lot, while low value means the feature was not observed much. Based only on this data, what explains their course-failure prediction, with the predictive grade of 3.5/6? Please explain week-by-week, in easy terms the feature names, as if you were explaining to a high-school student what impact these features have on their work, and why they are predicted to failing the course.

Feature: StudentSpeed_InWeek1, value of the feature: -0.0067532565444707, importance of the feature : 0.0317289155033689, the class average value : 8.43484829051161 Feature: StudentSpeed_InWeek2, value of the feature: -0.0055191592127084, importance of the feature : 0.0451485386922976, the class average value : 10.616412048968728 Feature: StudentSpeed_InWeek3, value of the feature: -0.0007102355360984, importance of the feature : 0.0333695532025385, the class average value : 9.480981508371652 Feature: StudentSpeed_InWeek4, value of the feature: -0.0098639556765556, importance of the feature : 0.0, the class average value : 5.481381946170677 Feature: StudentSpeed_InWeek5, value of the feature: -0.0090934773534536, importance of the feature : 0.1157640642619183, the class average value : 8.261572731376528 Feature: NumberSubmissions-perc_correct_InWeek1, value of the feature: 0.006664959937334, importance of the feature : 0.0, the class average value : 36.33383521432614 Feature: NumberSubmissions-perc_correct_InWeek2, value of the feature: 0.0028375887870788, importance of the feature : 0.2854288568574283, the class average value : 38.55844712373066 Feature: NumberSubmissions-perc_correct_InWeek3, value of the feature: 5.386002361773557e-05, importance of the feature : 0.5550005550005551, the class average value : 49.65379789729113 Feature: NumberSubmissions-perc_correct_InWeek4, value of the feature: 0.0113101581484079, importance of the feature : 0.0, the class average value : 51.86682691724709 Feature: NumberSubmissions-perc_correct_InWeek5, value of the feature: 0.0098912681639194, importance of the feature : 0.0, the class average value : 31.92182273013053 Feature: FrequencyEvent-video.load_InWeek1, value of the feature: 0.0057042513042688, importance of the feature : 0.3719357565511412, the class average value : 14.274109513243257 Feature: FrequencyEvent-video.load_InWeek2, value of the fea-

ture: 0.002497244849801, importance of the feature : 0.1696416790756413, the class average value : 9.277202098840547 Feature: FrequencyEvent-video.load_InWeek3, value of the feature: 0.0008539947867393, importance of the feature : 0.1064384626028461, the class average value : 10.50574398876291 Feature: FrequencyEvent-video.load_InWeek4, value of the feature: 0.0001481300592422, importance of the feature : 0.0832500832500832, the class average value : 9.618689460397677 Feature: FrequencyEvent-video.load_InWeek5, value of the feature: 0.013764046728611, importance of the feature : 0.5994005994005994, the class average value : 11.215158870359595 Feature: TimeSolveProblem_InWeek1, value of the feature: -0.0024096494168042, importance of the feature : 0.5487113937159338, the class average value : 25.221784707944238 Feature: TimeSolveProblem_InWeek2, value of the feature: -0.0032676710933446, importance of the feature : 0.5715950124358583, the class average value : 20.811933630013172 Feature: TimeSolveProblem_InWeek3, value of the feature: -0.0036968085914849, importance of the feature : 0.5374183937096526, the class average value : 15.519625810926962 Feature: TimeSolveProblem_InWeek4, value of the feature: -0.0087775432318449, importance of the feature : 0.9990002881731602, the class average value : 20.42804936522804 Feature: TimeSolveProblem_InWeek5, value of the feature: -0.002507562637329, importance of the feature : 0.4251432851446636, the class average value : 16.645763136034713 Feature: problem.check-problem.check-video.load_InWeek1, value of the feature: 0.0013302743434906, importance of the feature : 0.1665001665001665, the class average value : 23.144522144522146 Feature: problem.check-problem.check-video.load_InWeek2, value of the feature: 0.0040758438408374, importance of the feature : 0.1665001665001665, the class average value : 35.63203463203463 Feature: problem.check-problem.check-video.load_InWeek3, value of the feature: 0.0027376251667737, importance of the feature : 0.1762942939413527, the class average value : 26.621437386143263 Feature: problem.check-problem.check-video.load_InWeek4, value of the feature: 0.0069088946282863, importance of the feature : 0.0, the class average value : 32.39717425431711 Feature: problem.check-problem.check-video.load_InWeek5, value of the feature: 0.0039117003977298, importance of the feature : 0.0, the class average value : 20.506808980493194

ChatGPT:

Let's break down Charlie's mid-semester performance week by week based on the provided data:

Week 1:

- Student Speed: Charlie starts a bit slower than most students. It's like they're taking a little more time than others to get going on the work.
- Submissions and Correctness: They're submitting work, but it's not exceptional in terms of correctness. It's around the class average.
- Watching Videos: Charlie watches some videos, but not a lot compared to the rest of the class. Seems like an average start here.
- Time to Solve Problems: It takes Charlie a bit more time than most to solve problems, but it's not too far off the average.
- Combining Check Problems with Video Watching: They do some check problems related to watching videos but not too many. It's moderate compared to others.

Week 2:

- Student Speed: Charlie's speed in getting work done decreases a bit more compared to the first week. They're taking even more time now.
- Submissions and Correctness: The correctness of their submissions improves slightly but nothing extraordinary. They're around the same as the first week.
- Watching Videos: They watch a few more videos than the first week, trying to catch up or understand better, perhaps.
- Time to Solve Problems: Charlie takes a bit more time compared to before. The problems might be getting a tad harder for them.

- Combining Check Problems with Video Watching: They increase doing check problems related to watching videos. Maybe trying to understand the material better.

Week 3:

- Student Speed: Charlie's speed decreases a bit more. They're taking even longer now compared to previous weeks.
- Submissions and Correctness: The correctness of their submissions drops noticeably. It's getting tougher for Charlie to get things right.
- Watching Videos: They watch about the same as the second week, not much change here.
- Time to Solve Problems: Charlie's taking even more time now. Problems might be getting quite challenging.
- Combining Check Problems with Video Watching: They keep up with doing some check problems related to watching videos but not as much as before.

Week 4:

- Student Speed: Charlie's speed drops significantly. They're taking way longer than most students to complete work.
- Submissions and Correctness: Not much information on correctness, but their submissions might be slowing down significantly.
- Watching Videos: They watch way less than previous weeks. Maybe they're finding it harder to keep up or understand.
- Time to Solve Problems: Charlie takes even more time, significantly more than most classmates.
- Combining Check Problems with Video Watching: No information on this aspect.

Week 5:

- Student Speed: Charlie's speed remains quite slow, similar to the previous week.
- Submissions and Correctness: Again, not much detail on correctness, but it might be affected by their slowing pace.
- Watching Videos: They start watching more videos again, possibly trying to catch up in the last moments.
- Time to Solve Problems: Still taking a lot of time, indicating continued struggle.
- Combining Check Problems with Video Watching: No information provided for this week.

Summary : Charlie's slow engagement and limited time spent solving problems were the primary reasons for the predicted course failure. Although they showed improvement in correctness rates and maintained a good frequency of watching videos, these factors were not enough to offset the impact of their slower pace and reduced problem-solving time.

7.2 TEXT EXPLANATION 2 GENERATION

Here are the interactions with ChatGPT 3.5 to generate the Text Explanation 2 for student id 206.

User :

You are a talented data scientist and learning scientist. Bellow are some features, with their value and importance scores. You want to take the following data to generate a natural sounding, short, concise explanation of the ungendered student Charlie performance for a mid-semester intervention. Positive scores mean that the feature corresponded to student success, negative scores means that the feature contributed towards student failure. High value means the feature was observed a lot, while low value means the feature was not observed much. Based only on this data, what explains their course-failure prediction, with the predictive grade of 3.5/6? Please synthesis week-by-week, in easy terms the feature names, as if you were explaining to a high-school student what impact these features have on their work,

and why they are predicted to failing the course.

Prediction: failure, predictive grade: 3.5/6 Feature: StudentSpeed_InWeek1, value of the feature: -0.0067532565444707, importance of the feature : 0.0317289155033689 Feature: StudentSpeed_InWeek2, value of the feature: -0.0055191592127084, importance of the feature : 0.0451485386922976 Feature: StudentSpeed_InWeek3, value of the feature: -0.0007102355360984, importance of the feature : 0.0333695532025385 Feature: StudentSpeed_InWeek4, value of the feature: -0.0098639556765556, importance of the feature : 0.0 Feature: StudentSpeed_InWeek5, value of the feature: -0.0090934773534536, importance of the feature : 0.1157640642619183 Feature: NumberSubmissions-perc_correct_InWeek1, value of the feature: 0.006664959937334, importance of the feature : 0.0 Feature: NumberSubmissions-perc_correct_InWeek2, value of the feature: 0.0028375887870788, importance of the feature : 0.2854288568574283 Feature: NumberSubmissions-perc_correct_InWeek3, value of the feature: 5.386002361773557e-05, importance of the feature : 0.5550005550005551 Feature: NumberSubmissions-perc_correct_InWeek4, value of the feature: 0.0113101581484079, importance of the feature : 0.0 Feature: NumberSubmissions-perc_correct_InWeek5, value of the feature: 0.0098912681639194, importance of the feature : 0.0 Feature: FrequencyEvent-video.load_InWeek1, value of the feature: 0.0057042513042688, importance of the feature : 0.3719357565511412 Feature: FrequencyEvent-video.load_InWeek2, value of the feature: 0.002497244849801, importance of the feature : 0.1696416790756413 Feature: FrequencyEvent-video.load_InWeek3, value of the feature: 0.0008539947867393, importance of the feature : 0.1064384626028461 Feature: FrequencyEvent-video.load_InWeek4, value of the feature: 0.0001481300592422, importance of the feature : 0.0832500832500832 Feature: FrequencyEvent-video.load_InWeek5, value of the feature: 0.013764046728611, importance of the feature : 0.5994005994005994 Feature: TimeSolveProblem_InWeek1, value of the feature: -0.0024096494168042, importance of the feature : 0.5487113937159338 Feature: TimeSolveProblem_InWeek2, value of the feature: -0.0032676710933446, importance of the feature : 0.5715950124358583 Feature: TimeSolveProblem_InWeek3, value of the feature: -0.0036968085914849, importance of the feature : 0.5374183937096526 Feature: TimeSolveProblem_InWeek4, value of the feature: -0.0087775432318449, importance of the feature : 0.9990002881731602 Feature: TimeSolveProblem_InWeek5, value of the feature: -0.002507562637329, importance of the feature : 0.4251432851446636 Feature: problem.check-problem.check-video.load_InWeek1, value of the feature: 0.0013302743434906, importance of the feature : 0.1665001665001665 Feature: problem.check-problem.check-video.load_InWeek2, value of the feature: 0.0040758438408374, importance of the feature : 0.1665001665001665 Feature: problem.check-problem.check-video.load_InWeek3, value of the feature: 0.0027376251667737, importance of the feature : 0.1762942939413527 Feature: problem.check-problem.check-video.load_InWeek4, value of the feature: 0.0069088946282863, importance of the feature : 0.0 Feature: problem.check-problem.check-video.load_InWeek5, value of the feature: 0.0039117003977298, importance of the feature : 0.0

ChatGPT:

Let's break down Charlie's mid-semester performance week by week based on the provided data:

Week 1: Charlie started off a bit slow in understanding the course material. They took more time than usual to work through the lessons, which might have made it harder to keep up. Also, their accuracy in solving problems wasn't quite where it needed to be.

Week 2: Similar to the first week, Charlie continued to work through the material at a slower pace. They still struggled a bit with getting the right answers in the assignments.

Week 3: Charlie's pace improved slightly compared to the previous weeks but was still slower than what's typically expected. Their accuracy in solving problems showed a bit of improvement as well.

Week 4: There's a significant drop here. Charlie's speed in going through the course material slowed down again, and their accuracy in solving problems took a hit.

Week 5: Towards the end, Charlie tried to catch up by watching more course videos, which is good! However, they were still spending a lot of time on problem-solving tasks, which might mean they were finding it challenging to understand and solve the course problems efficiently.

Summary: Charlie struggled with maintaining a good pace in learning, faced difficulties in accuracy, and although they tried to catch up by watching more videos, they still took a lot of time in solving problems. This combination likely affected their overall performance in the course.

7.3 TEXT EXPLANATION 3 GENERATION

Here are the interactions with ChatGPT 3.5 to generate the Text Explanation 3 for student id 206.

User :

You are a talented data scientist and learning scientist. Bellow are some features, with their value and importance scores. You want to take the following data to generate a natural sounding, short, concise explanation of the ungendered student Charlie performance for a mid-semester intervention. Positive scores mean that the feature corresponded to student success, negative scores means that the feature contributed towards student failure. High value means the feature was observed a lot, while low value means the feature was not observed much. Based only on this data, what explains their course-failure prediction, with the predictive grade of 3.5/6? Please synthesis feature-by-feature, in easy terms the feature names, as if you were explaining to a high-school student what impact these features have on their work, and why they are predicted to failing the course.

Prediction: failure, predictive grade: 3.5/6 Feature: StudentSpeed_InWeek1, value of the feature: -0.0067532565444707, importance of the feature : 0.0317289155033689 Feature: StudentSpeed_InWeek2, value of the feature: -0.0055191592127084, importance of the feature : 0.0451485386922976 Feature: StudentSpeed_InWeek3, value of the feature: -0.0007102355360984, importance of the feature : 0.0333695532025385 Feature: StudentSpeed_InWeek4, value of the feature: -0.0098639556765556, importance of the feature : 0.0 Feature: StudentSpeed_InWeek5, value of the feature: -0.0090934773534536, importance of the feature : 0.1157640642619183 Feature: NumberSubmissions-perc_correct_InWeek1, value of the feature: 0.006664959937334, importance of the feature : 0.0 Feature: NumberSubmissions-perc_correct_InWeek2, value of the feature: 0.0028375887870788, importance of the feature : 0.2854288568574283 Feature: NumberSubmissions-perc_correct_InWeek3, value of the feature: 5.386002361773557e-05, importance of the feature : 0.5550005550005551 Feature: NumberSubmissions-perc_correct_InWeek4, value of the feature: 0.0113101581484079, importance of the feature : 0.0 Feature: NumberSubmissions-perc_correct_InWeek5, value of the feature: 0.0098912681639194, importance of the feature : 0.0 Feature: FrequencyEvent-video.load_InWeek1, value of the feature: 0.0057042513042688, importance of the feature : 0.3719357565511412 Feature: FrequencyEvent-video.load_InWeek2, value of the feature: 0.002497244849801, importance of the feature : 0.1696416790756413 Feature: FrequencyEvent-video.load_InWeek3, value of the feature: 0.0008539947867393, importance of the feature : 0.1064384626028461 Feature: FrequencyEvent-video.load_InWeek4, value of the feature: 0.0001481300592422, importance of the feature : 0.0832500832500832 Feature: FrequencyEvent-video.load_InWeek5, value of the feature: 0.013764046728611, importance of the feature : 0.5994005994005994 Feature: TimeSolveProblem_InWeek1, value of the feature: -0.0024096494168042, importance of the feature : 0.5487113937159338 Feature: TimeSolveProblem_InWeek2, value of the feature: -0.0032676710933446, importance of the feature : 0.5715950124358583 Feature: TimeSolveProblem_InWeek3, value of the feature: -0.0036968085914849, importance of the feature : 0.5374183937096526 Feature: TimeSolveProblem_InWeek4, value of the feature: -0.0087775432318449, importance of the feature : 0.9990002881731602 Feature: TimeSolveProblem_InWeek5, value of the feature: -0.002507562637329, importance of the feature : 0.4251432851446636 Feature: problem.check-problem.check-video.load_InWeek1, value of the

feature: 0.0013302743434906, importance of the feature : 0.1665001665001665 Feature: problem.check-problem.check-video.load_InWeek2, value of the feature: 0.0040758438408374, importance of the feature : 0.1665001665001665 Feature: problem.check-problem.check-video.load_InWeek3, value of the feature: 0.0027376251667737, importance of the feature : 0.1762942939413527 Feature: problem.check-problem.check-video.load_InWeek4, value of the feature: 0.0069088946282863, importance of the feature : 0.0 Feature: problem.check-problem.check-video.load_InWeek5, value of the feature: 0.0039117003977298, importance of the feature : 0.0

ChatGPT:

Let's break down Charlie's mid-semester performance for each feature based on the provided data:

Student Speed in Each Week: Charlie has been moving through the course material slower than most other students throughout the weeks. This might mean they're taking more time than usual to understand the topics.

Accuracy in Problem Solving: Charlie started off decently in the second and third weeks but struggled in the subsequent weeks, consistently getting fewer correct answers than their classmates. This suggests a decline in their understanding or application of the concepts.

Frequency of Watching Videos: Charlie has been watching a lot of course videos, especially in the last week. While it shows effort in learning, it might indicate they're spending more time watching rather than practicing problem-solving.

Time Taken to Solve Problems: Across all weeks, Charlie has been consistently taking a longer time to solve problems compared to others. This suggests that they find the coursework challenging and might need extra help or practice to improve.

Summary: Taking these factors together, Charlie seems to be having difficulty keeping up with the course pace, understanding the material thoroughly, and applying it effectively within the given time. This combination might be contributing to the prediction of course failure. However, there's potential for improvement if Charlie can manage their time better and focus on practicing problem-solving techniques.

BIBLIOGRAPHY

- Swamy, Vinitra, Bahar Radmehr et al. (2022). *Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs*. arXiv: 2207.00551 [cs.LG].
- Haque, AKM Bahalul, A.K.M. Najmul Islam and Patrick Mikalef (2023). ‘Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research’. In: *Technological Forecasting and Social Change* 186, p. 122120. ISSN: 0040-1625. DOI: <https://doi.org/10.1016/j.techfore.2022.122120>. URL: <https://www.sciencedirect.com/science/article/pii/S0040162522006412>.
- Swamy, Vinitra, Jibril Frej and Tanja Käser (2023). *The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations*. arXiv: 2307.00364 [cs.LG].
- Hostetter, John Wesley et al. (2023). ‘XAI to Increase the Effectiveness of an Intelligent Pedagogical Agent’. In: *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. IVA ’23. <conf-loc>, <city>Würzburg</city>, <country>Germany</country>, </conf-loc>: Association for Computing Machinery. ISBN: 9781450399944. DOI: 10.1145/3570945.3607301. URL: <https://doi.org/10.1145/3570945.3607301>.
- Rachha, Ashwin and Mohammed Seyam (2023). ‘Explainable AI In Education : Current Trends, Challenges, And Opportunities’. In: *SoutheastCon 2023*, pp. 232–239. DOI: 10.1109/SoutheastCon51012.2023.10115140.
- Vinitra Swamy Juliette Parchet, Tanya Nazaretsky (2023). *ML4ED Robutness of XAI study interviewees*. <https://drive.google.com/drive/folders/1EZqG5S5bwgp9yRZ-79CZE7Xe21Mlqqvh?usp=sharing>.
- Nazaretsky, Tanya et al. (2022). ‘Teachers’ trust in AI-powered educational technology and a professional development program to improve it’. In: *British journal of educational technology* 53.4, pp. 914–931.
- Parchet, Juliette (2023). *Robutness of XAI*. https://github.com/Jucifer06/ML4Ed_semester_project.git.
- Meloncon, Lisa and Emily Warner (July 2017). ‘Data visualizations: A literature review and opportunities for technical and professional communication’. In: pp. 1–9. DOI: 10.1109/IPCC.2017.8013960.
- OpenAI (Dec. 2023). *ChatGPT 3.5*. <https://chat.openai.com/>.
- Juliette Parchet Vinitra Swamy, Tanya Nazaretsky (2023). *Providing explanations of student behavior for TA / Teacher intervention design*. <https://drive.google.com/drive/folders/1-DuQ5EoSU3wgQ3LzKLHWPuYQwqCyJN1U?usp=sharing>.