Can Students Understand AI Decisions Based on Variables Extracted via AutoML?

Liang Tang
School of Information Sciences
University of Illinois Urbana–Champaign
Champaign, USA
ltang29@illinois.edu

Nigel Bosch
School of Information Sciences and Department of Educational Psychology
University of Illinois Urbana–Champaign
Champaign, USA
pnb@illinois.edu

***Abstract*—In computer-based education, understanding student data is essential for students, teachers, researchers, and others to adapt to insights gained from analyses (e.g., AI predictions of student outcomes). However, one important question is: how well can students make sense of the data we present? And what factors influence the interpretability of those data? This study assessed students' perceptions of predictive variables (i.e., "features") used in machine learning models for predicting student outcomes; in particular, we explored features crafted by experts versus those extracted by methods for automatic machine learning (i.e., AutoML). Our results indicated a meaningful difference in students' interpretability perceptions between the expert and AutoML features across two diverse datasets. Additionally, features derived from timing and scoring data were more interpretable than those from interaction (e.g., keystroke) data. Other potential explanations for interpretability differences, including statistical methods, repeated exposure, and lexical familiarity, had relatively minimal impact on interpretability.**

I. Introduction

The shift towards e-learning and web-based education methods has led to promising opportunities to expand and improve education [34]. For example, within the context of computer-mediated learning, the abundant data generated from teacher–student interactions can be leveraged to optimize the digital interface for both parties [2]. Machine learning is becoming a transformative tool not only for institutions of higher education but, more importantly, for enhancing student success directly [38]. Machine learning promises to provide insights that can reshape academic processes and interventions and has been used to systematically explore complex educational datasets to identify hidden patterns and relationships [43]. Through insights derived from data mining, educators have the opportunity to refine their curriculum in alignment with students' needs, promoting engagement and optimizing learning outcomes [26], [44]. Importantly, this process can also empower students by providing them access to insights mined from their own learning data, fostering a more personalized and self-directed learning experience. As students engage with their data, they gain valuable insights into their learning habits, strengths, and areas for improvement [15]. Many scholarly works have explored the utilization of machine learning techniques in this domain, concentrating on collecting student data, analyzing their behavior, and identifying—for example—students who might encounter academic difficulties [33], [42]. Such advancements have potential to radically transform outcomes, from predicting student performance to optimizing curriculum delivery. However, the introduction of such technology also carries with it ethical, practical, and legal responsibilities to ensure that the decisions from machine learning models can be understood—especially when stakeholders are directly affected by predictions or are expected to learn from them, as may often be the case with students.

Historically, the primary focus of machine learning research was on performance optimization, with the aim of designing systems capable of delivering the most accurate predictions by, for example, learning rules and decision trees from data [28]. As AI models became more accurate, a challenge emerged: these advanced predictive models were often too complex to understand, leaving researchers and students puzzled about the "how" and "why" behind models' predictions [14]. This realization has prompted a rapid increase in research focusing on the interpretability of AI systems [12], and especially on designing methods to enable stakeholders to explore what inputs are most relevant to a prediction and how particular values affect a prediction [19]. However, a critical yet poorly understood consideration for interpreting machine learning model decisions arises from the fact that interpreting a model is intrinsically tied to the inputs, or *features*, that it uses. Notably, even a simple linear regression model can become incomprehensible when trained with uninterpretable features.

Recognizing students as active users of machine learning-driven tools is a critical shift in perspective. These tools, such as student-facing dashboards, allow students to understand and reflect upon the predictions made about their learning outcomes [37]. By accessing and interpreting their learning data, students transition from being mere subjects of educational analysis to becoming interactive users, empowering them to take control of their educational process and change their learning strategies accordingly. The impact of dashboard learning analytics systems on students' learning experiences is profound. Students engage more deeply with their education when they can interpret the data presented to them [23]. Given the critical role that students play as users, the interpretability of data used for machine learning in education is a pedagogical necessity beyond a mere technical concern. Students have the opportunity to self-reflect and learn from machine learning

models if they can interpret, for instance, *why* a model makes a prediction about a particular learning outcome and how they can adjust accordingly [5], [20]. Enabling students to understand and reflect upon the predictions made about their learning outcomes is fundamental, ensuring that educational technology enhances the overall learning experience by being meaningful to all stakeholders, especially students. Overall, the primary objective of this work is to explore how students can understand machine-generated features in educational machine-learning models to some extent. By investigating the ways in which students can interpret and engage with these features, we aim to bridge the gap between the technical aspects of machine learning and the pedagogical needs of students.

II. Background

A. Interpretability

The concept of interpretability is crucial in developing understandable machine learning models, referring to humans' ability to comprehend the reasoning behind a model's decision [8]. Machine learning interpretability is important for several reasons, such as enabling experts to refine models [27], fostering trust [6], and ensuring that students and teachers can comprehend, learn from, and question a model's evaluations in educational scenarios [7], [20]. Studies reveal that when students can interpret and engage with models predicting their learning outcomes, they can better adjust their learning strategies, leading to improved educational experiences and higher acceptance [6]. Interpretability also plays a role in addressing societal concerns, such as the European Union's establishment of rights for individuals to access meaningful information about the logic involved in automated decision-making [39]. Emphasizing the "why" behind AI decisions is a fundamental feature of responsible AI deployment [17]. A student's ability to interpret a model is influenced by various factors, with features being a particularly predominant influence [31]. Considering interpretability implies that there must be someone to do the interpreting, and different stakeholders may require various levels and forms of feature interpretability. In this paper, we focus on students, the largest and most directly impacted group of stakeholders in education. We expect that the effectiveness of AutoML tools in educational contexts critically depends on the degree to which students can understand and interact with these systems. Education systems designed with student interpretability in mind act not just as tools for assessment but as catalysts for active learning and self-reflection. In this study, we define interpretability as the extent to which students can understand and make sense of the features used in machine learning models for predicting educational outcomes.

B. Feature Engineering for Machine Learning with Educational Data

As the volume of educational data expands, the insights we seek to extract become more complex [10]. Designing and extracting features (i.e., *feature engineering*) from complex learning data in a manner that is readily digestible for students and teachers poses a challenge [22], yet an important one since ML models are only as interpretable as their features [45]. Therefore, ensuring that these features are both meaningful and interpretable to educators and students is critical for the effective use of ML in education.

Automated machine learning (AutoML) methods have emerged to address the challenge of feature engineering, offering greater adaptability and efficiency compared to manual feature selection. We focus specifically on AutoML for the feature engineering step [9], which can affect the interpretability of any type of model or training process. Tools like FeatureTools [3] (for relational data) and TSFRESH [11] (Time Series Feature Extraction based on Scalable Hypothesis tests) automate the feature engineering process by extracting huge numbers of features based on provided data and metadata. Expert-driven feature engineering methods rely on domain knowledge and human intuition, leading to features that we expect to be more interpretable and contextually relevant, but potentially limited in scope due to cognitive biases and time/effort required. In contrast, AutoML systems utilize mathematical and statistical methods, allowing for a more comprehensive and unbiased exploration of data. However, understanding their interpretability is still an unknown challenge. Previous work indicates that domain experts find AutoML features more difficult to interpret [3], but little is known about students' perceptions.

C. The Present Study

From a student's perspective, the interpretability of a system might not always directly influence their decision to accept it. Instead, interpretability might more subtly impact human confidence level, trust, comfort, and acceptance with the technology [6], [35], thereby impacting their acceptance of utilizing these technologies in their learning process. For example, when the visual representation of data and personalized performance feedback work together to build user trust, users are more likely to trust the machine learning result because they can clearly see how their data are being used [1]. Nonetheless, if the learning analytics dashboard's inner workings are opaque, the student might use it with a degree of skepticism or discomfort, which could eventually affect their continued engagement with the dashboard.

Researchers have increasingly concentrated on exploring the impact of machine learning features' characteristics (such as perceived usefulness, complexity, performance expectancy, and effort expectancy) [21]. One notable area of research is the concept of "procedural fairness" in algorithmic decision-making [13], which concerns how people perceive the fairness of outcomes delivered by algorithms. This is closely tied to the features themselves, as perceived fairness can be influenced by how users understand the features that lead to predicted outcomes.

Traditionally, educational machine learning research has focused on the accuracy and efficiency of predictive models [4], with less emphasis on interpretability from the perspective of non-technical end-users like students. Our study evaluates

the interpretability of machine learning features, particularly those generated by AutoML methods, addressing a gap in the existing literature. We investigate which features are more interpretable for students and why, aiming to make AutoML methods more accessible and understandable in educational contexts. We address two research questions:

> *RQ1: In the context of predicting student outcomes, how does the interpretability of AutoML features compare to those crafted by experts?*

Hypothesis: AutoML features will be harder to grasp than expert-developed features because domain experts are guided by similar educational experiences and intuitions as students, while AutoML methods lack biases toward domain requirements or interpretability preferences.

> *RQ2: Apart from their origin (i.e., AutoML vs. expert), what characteristics make a feature more interpretable to students?*

Hypothesis: The intricacy of statistical methods, aggregation functions, and familiarity with terms used might influence a feature's interpretability [32].

III. Method

A. Machine Learning Data Description

This paper relies on two datasets from which we extracted features and trained predictive models for students to evaluate. The Open University Learning Analytics Dataset (OULAD) [24], collected between 2013 and 2014, captures student behaviors, interactions with course materials, and their final outcomes. OULAD consists of relational data tables, including student-level, student × assessment-level, and day-level data for 32,592 students and spans over 10 million rows describing students' interactions with an online learning management system. For modeling purposes, we converted the original four student outcome categories ("fail", "withdrawn", "pass", and "distinction") into binary categories (i.e., "fail" and "pass") to simplify the interpretation of the machine learning models' results and ensure consistency across different datasets.

The other dataset, Educational Process Mining (EPM) [40], records interface interaction data at a finer granularity than OULAD; EPM has data from 115 first-year undergraduate engineering students with 230,318 rows of data about their interactions with electronic tutoring software and assessment scores corresponding to different topics within the system. Student outcomes ranged from 0 to 5 and were used directly in prediction tasks without conversion.

The datasets both underwent a cleaning process that involved removing missing values and columns with no variance.

B. AutoML Feature Engineering

The differences in complexity and abstraction level between AutoML-generated and expert-crafted features are significant, stemming from their origin and the nature of how they are created. Even when translated into simpler terms, the underlying complexity and relationships they represent might not be immediately obvious. For instance, an AutoML system might use the first digits of a submission date as a feature, which abstractly connects the submission date with performance without a clear educational rationale, whereas a human expert might design a "beginning of the month" or "end of month" feature instead.

We applied FeatureTools [18] to the OULAD dataset and TSFRESH to the EPM dataset. FeatureTools with default settings yielded 3,799 features, which were reduced to 668 after removing features with no variance and redundant features. TSFRESH aggregates time series data, producing simple and complex features that may capture trends like student pacing or time allocation strategies (see section III-D). After eliminating invariant and redundant features, 321 of the original 6,312 features were retained. The choice to apply different AutoML methods was influenced by the unique characteristics and specific requirements of each dataset. TSFRESH specializes in extracting relevant features from time series data, making it suitable for the EPM dataset, while FeatureTools excels in dealing with relational datasets, making it more suitable for OULAD.

C. Expert Feature Engineering

We manually engineered features expected to be good predictors of students' assessment scores based on our expertise in educational technology, data mining, and machine learning, with substantial prior experience in feature engineering [Citations removed for anonymous review]. We prioritized predictive utility over interpretability and used preexisting features from prior research unrelated to interpretability [Citations removed for anonymous review] to reduce potential biases. These features represent common feature engineering work without emphasis on interpretability, serving as a "typical" comparison for AutoML methods.

The final dataset included 75 features for EPM and 28 for OULAD, with structure and design similar to other educational data mining literature [16], [25]. For example, we transformed quiz score data into (i) quantiles, (ii) indicators of exceeding class average, and (iii) standard deviations. We also combined *date submitted* and *date* features to form *past due*, referring to how often students submitted assignments past the due date.

D. Per-feature Analysis and Feature Selection

Given that there were hundreds of features for students to evaluate, we narrowed the list of features to those that were most useful for predicting student assessment scores. To do so, we trained a single decision tree model with 5-fold cross-validation via *scikit-learn* in Python [29]. We then ranked features according to accuracy and selected the top 15 for inclusion in a subsequent interpretability survey given to students. The accuracy metrics used for this evaluation were R^2 for the EPM data (a continuous assessment score outcome) and area under the receiver operating characteristic curve (AUC) for OULAD (a pass/fail outcome).

Our primary goal was to evaluate the comprehensibility of the fundamental concepts represented in the features, not

the internal feature names generated by AutoML (or even experts), which can be unnecessarily difficult to understand (e.g., "f_agg" may be hard to understand without knowing that TSFRESH uses this to denote using a function to aggregate a sequence). Hence, we manually "translated" the feature names into a more interpretable form. For example:

FeatureTools: MAX(assessmentsmerged.CUM_COUNT(assessment_type))
Translation: Largest value of the count of how many assignments were submitted by the student so far

TSFRESH: mouse_click_left_agg_linear_trend_attr_ "rvalue"__chunk_len_5__f_agg_"mean"
Translation: Correlation of a line drawn through the sequence of values that consist of the average number of mouse clicks in the last 5 actions done by the student

Students participating in the interpretability survey were exposed solely to translated descriptions of features. The key point of comparison in our study is how these features, originating from different sources, are perceived by students. This comparison is vital because, in real-world applications, feature names generated purely by machine learning algorithms might often be impractical due to their inherent complexity or lack of clarity. Such features, while mathematically or statistically well-motivated, may not effectively communicate the practical implications or relevance of the features they represent. Hence, the translated feature names represent a more realistic scenario of how features might be presented to students (e.g., in a student-facing dashboard).

E. Constructing Machine Learning Models

As described below in section III-F, the interpretability survey displayed predicted outcomes from machine learning models to student participants. To create these predictions, we trained and applied random forest classification and regression models using the AutoML-generated features and the hand-crafted features separately, with data split into testing and training sets using a 30/70 ratio. The accuracy of each model was assessed using standard metrics including AUC and R^2 (Table I).

Table I
MODEL ACCURACIES FOR MACHINE LEARNING MODELS USED TO PREDICT STUDENT OUTCOMES.

Dataset	Feature Type	Metric	Result
EPM	TSFRESH	R^2	.471
EPM	Expert	R^2	.443
OULAD	FeatureTools	AUC	.812
OULAD	Expert	AUC	.798

In our study, the decision to train a random forest model is driven by considerations that directly support our research objectives. Firstly, random forest was chosen for its well-established reputation and widespread use in machine learning, including within the field of education [36], [43]. Random forest's robustness against overfitting makes it an appropriate choice for prediction tasks with many features, such as in this study. Secondly, and more importantly, our research primarily concentrates on the analysis and selection of features, with the choice of model serving merely as a means to facilitate this focus; hence, comparisons between different model types would be largely orthogonal to the research questions in this study. The focus was not on comparing model types but rather on highlighting the utility of the features within typical EDM contexts. By prioritizing feature selection, we aim to ensure that our study's findings are grounded in the quality and relevance of features, rather than the capabilities of any particular modeling technique.

F. Assessing Interpretability with Students

We assessed interpretability by surveying university/college students across the United States (study approved by university institutional review board, protocol #[masked for review]).

We recruited college students via Prolific, an online data collection platform with flexible participant selection criteria and empirical evidence demonstrating its quality [30].

After consenting to participate, students viewed survey instructions and answered questions about their backgrounds, their experience with machine learning, and their understanding of both AutoML-generated and expert-crafted features. We integrated prediction tasks in the survey to help ground participants' interpretability judgments in a specific task. We opted to use a direct 1-5 scale for evaluation due to the number of features and the impracticality of applying a comprehensive survey for each one.

Additionally, previous research has shown that stakes and scarcity can influence positive attitudes towards AI interpretability [28], we incorporated stakes and scarcity as an element of the survey by revealing only specific resources (i.e., features) for decision-making, then allowing participants to view additional features if they could not decide. Participants received points based on prediction accuracy and whether they asked for additional features, as an incentive to attend to the task and interpret the features. Participants received $5 USD for completing the study, with those scoring in the top 10% receiving an additional $2.

We used the models from Table I as illustrations of machine learning predictions with student data during the survey. Based on Table I, we indicated to participants that "the algorithm has previously been evaluated on a large dataset of student data. The system has been found to be reliable, achieving a task accuracy of 75-85%, or performing 40-50% better than random guessing of a student's grade". In binary classification tasks, where random guessing would be correct 50% of the time (the baseline rate), saying a model is "40% better than random guessing" indicates that the model's performance exceeds this baseline rate by 40% of the possible improvement on

a 0–100% scale. The reason for expressing model performance in this manner is to provide a standardized way of comparing model effectiveness across different metrics and datasets. In our study, different models were evaluated using various metrics (i.e., R^2 and AUC), depending on what was most appropriate for the specific dataset. By relating model performance to the baseline of random guessing, we offer this common ground for comparison to participants without explaining the nuances of R^2 versus AUC interpretation.

Following the introduction, participants completed a set of prediction tasks. For each task, participants were shown a list of features derived from student activity, either by experts or AutoML, without an indication of which was the source. We provided the natural-language translation of feature names (per section III-D) and the values of the features for the given prediction task. Participants then made a prediction based on the features shown, then once more if they requested to see additional features (specifically, three additional features were shown). We then revealed the actual outcomes to help participants reflect and learn if they wished to improve. The tasks included five predictions from each of two datasets (EPM and OULAD) and two types of features (AutoML and expert), for a total of $5 \times 2 \times 2 = 20$ tasks, shown in randomized order to account for possible ordering effects and learning effects. After the final predictions were made, participants are asked to rate the interpretability of each encountered variable. We inquired about their understanding of each feature, using a 5-point scale where 1 signified "Not at all interpretable", 3 signified "Moderately interpretable", and 5 signified "Extremely interpretable".

IV. Results

A. Participants

In total, 199 college students participated from postsecondary institutions across the United States. Mean age was approximately 29.6 years, with a range from 18 to 74 years. The majority (53.5%) identified as male, 45.0% as female, and 8.50% as genders not reported in detail due to small sample sizes. Additionally, 61.3% identified as White, 15.6% as Asian, 13.1% as Black, and 10.0% as multiracial and additional races or ethnicities. English was the primary language spoken by 92.5% of participants. All participants were students and resided in the United States, with 41.5% employed full-time and 29.0% part-time. The sample primarily represented students with a limited range of expertise levels in AI/ML, from novices to those with intermediate knowledge.

The survey data and code have been made publicly available on the Open Science Framework at https://osf.io/nxguq/?view_only=62ff3fd8469c42d29aff19eb6d2cec4d

B. RQ1: Interpretability of Expert- versus AutoML-Generated Features

The results of the *Kruskal–Wallis* test (*H-value* = 315.607, $p < .001$) indicated that expert-crafted features were perceived as significantly more interpretable than AutoML features (Figure 1).

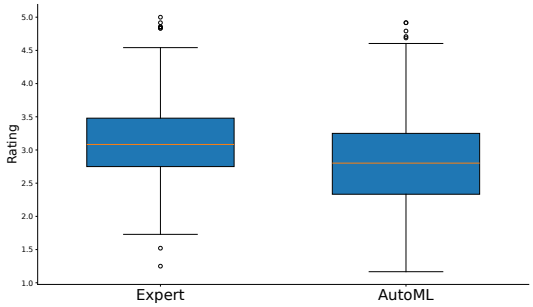

Figure 1. Box plot of average interpretability of AutoML versus expert-crafted features

It is important to note that the interpretability rating data in this study were ordinal. Ordinal data, by nature, are not expected to be normally distributed as they represent categories with a meaningful order but without consistent distances between categories. Thus, we applied non-parametric statistics.

C. RQ2: Attributes of Features Related to Interpretability

Given that RQ1 results showed significant differences between AutoML and expert features, despite representing both feature types in natural language, in RQ2 we further investigated what might explain these differences. We investigated potential sources of differences including the underlying type of data represented by a feature (section IV-C1), the aggregation functions applied in a feature (section IV-C2), and the influence of familiarity (section IV-C3).

1) Interpretability of Features' "Root" Data: We tested whether the original column(s) of data, or the "root" from which a feature was extracted, related to feature interpretability. For example, might prior assessment scores be considered more interpretable than lower-level data like the number of keystrokes a student does? We counted the appearance of each root column name, then filtered out those that appeared fewer than five times in the features shown to students (i.e., the most predictive features), yielding a set of five common roots. Then, we calculated the average interpretability rating of all features including that root.

From our investigation of root columns and their impact on interpretability, distinct patterns emerged (Figure 2). Perhaps unsurprisingly, students perceived score-based features (which appeared 8 times in the survey) as most interpretable for outcome prediction. Conversely, the *keystroke* "root", with the highest appearance count of 13, alongside its moderate interpretability rating of 2.66, suggests that while keystrokes are a common part of the learning environment and a good predictor of outcome, their significance or utility might not be fully clear to students. Similarly, *time* and *click* roots, appearing 12 and 11 times respectively, received moderate interpretability scores of 2.91 and 2.68. Another root similar to *time* is the *date* root, with an interpretability score of 2.98 (and 8 appearances). The close interpretability scores of *time* and *date* suggest that students have a comparable, relatively high level of understanding of both features. *mousewheel*

interactions, despite being the least interpretable root, had a substantial presence in the dataset with 10 appearances, indicating that mousewheel-based features were among the most effective for prediction.

A Kruskal–Wallis test further substantiated these observations (*H-value* = 221.235, $p < .001$), revealing a statistically significant difference in interpretability ratings among the roots.

2) Interpretability of Aggregation Functions: We investigated both the number of aggregation functions applied in feature extraction and the type of functions as explanations for differences in interpretability.

Number of Aggregation Functions. Features may consist of data aggregated at multiple levels; for example, a feature might simply be a mean of numeric values (one level), or the standard deviation of means of chunks of data (two levels), and so on. We explored aggregation up to three levels, where level 3 included 3 or more, expecting that more levels of aggregation would be perceived as less interpretable. We conducted this analysis in the AutoML features only because expert-engineered features did not tend to include more than one level of aggregation.

A Kruskal–Wallis test revealed a small significant difference between the three levels of aggregation (*H-value* = 82.016, $p < .001$). However, the levels were not entirely ordered as expected; mean interpretability for the three levels was 2.73, 2.93, and 2.59, respectively. Individual results per AutoML method (FeatureTools or TSFRESH) were also significant, but not in increasing order. For the FeatureTools method dataset (*H-value* = 26.432, $p < .001$), mean interpretability ratings were 2.97, 3.27, and 2.81. For TSFRESH (*H-value* = 24.660, $p < .001$), mean interpretability ratings were 2.52, 2.35, and 2.56 for the three layers respectively.

Type of Aggregation Functions. We averaged the interpretability ratings for features containing each statistical aggregation function, for functions that occurred at least five times. A Kruskal–Wallis revealed significant differences across aggregation functions ($F = 38.126$, $p = .004$). Figure 3 shows the results for 12 common aggregation functions (appearance > 5), suggesting that functions involving cumulative and proportional calculations, followed by those identifying the maximum values (max/highest) were considered most interpretable by participants.

3) Interpretability versus Feature Familiarity: **Recurrent Exposure.** We expected that repeated exposure to an aggregation function might increase interpretability due to higher perceived familiarity, and thus calculated a correlation between interpretability ratings and aggregation function frequency (i.e., how often participants encountered a particular aggregation function during the survey). However, Spearman's correlation results were not significant (*rho* = .009, $p = .983$), indicating no support for this hypothesis.

Interpretability versus Lexical Familiarity. We also expected that interpretability might be affected by lexical familiarity. The Brown Corpus [41] is a well-established corpus in linguistic studies, known for its broad representation of the English language across various genres and topics. By using word frequencies from this corpus, we aimed to obtain a reliable measure of lexical familiarity that is not limited to a specific domain or context. For example, perhaps binarization in Figure 3 is considered less interpretable than a proportion because "binary" is a less common/familiar word (and concept) than "proportion". We extracted word frequencies from the Brown corpus [41], removed stop words, then used the frequencies as an index of familiarity for words in the feature descriptions shown to participants, excluding words that appeared fewer than five times. However, the Spearman's correlation between word frequency and familiarity was not significant (*rho* = -.029, $p = .837$), so we could not conclusively state that familiarity with words has a direct relationship.

V. Discussion

In this study, we sought to understand how students assess the interpretability of features used for machine learning decisions in educational contexts. The distinction between expert and AutoML-generated features holds broad implications for contexts where interpretability is paramount, especially in education. While AutoML is convenient and accurate, it may not

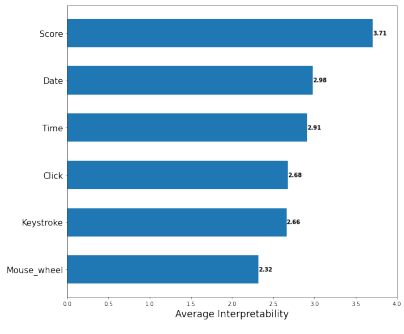

Figure 2. Interpretability ratings averaged by the root data from which features were derived.

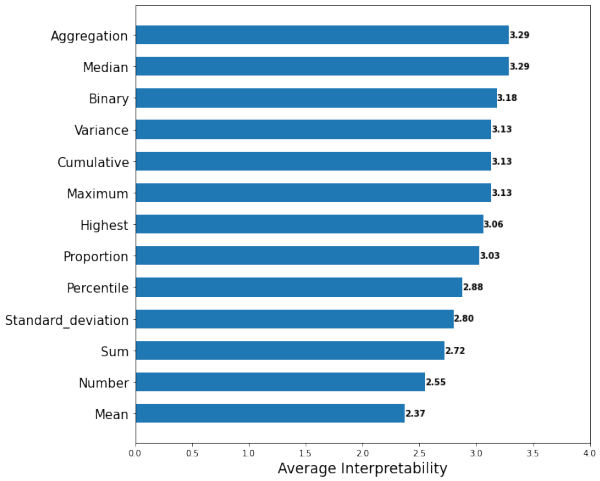

Figure 3. Interpretability of aggregation functions.

always align with human-centric goals like explainability. As RQ1 results showed, features crafted by human experts were, on average, significantly more interpretable despite efforts to make both expert and AutoML features as understandable as possible.

A. Factors Impacting Interpretability

Our study assessed potential factors influencing the interpretability of machine learning features in an educational context, including repeated exposure, lexical familiarity, the source or "root" data before aggregation, and type of aggregation in the features. Surprisingly, our findings did not show significant effects for repeated exposure and lexical familiarity, in contrast to our hypothesis for RQ2.

However, we observed significant differences in interpretability based on the root data and aggregation functions used for a feature, which has implications for interpretable adaptive learning systems. Interpretable adaptive learning systems will likely enjoy higher interpretability with features related to scores and timing, as opposed to interaction activity data (Figure 2). Certain aggregation functions may be preferable over others as well (Figure 3), though the reason behind these differences remains to be discovered. Additionally, as features undergo multiple layers of aggregation, their original context might become obfuscated, leading to a loss in clarity and interpretability. As we found (section IV-C2), the number of aggregation levels was related to interpretability, though not in the expected monotonically increasing way; hence, future work is needed to better understand this result and its implications for designing interpretable machine learning models for education.

B. Limitations and Future Work

There are a few opportunities for future work to address the limitations of this study. Since our current scope was limited to short time frame assessment, it remains unknown what the lasting impacts on interpretability might be, especially if students have time to learn more about the features over the course of a semester. Additionally, we used incentive rewards via virtual points as the sole criterion for motivating students, which may result in different goals for students to interpret features versus goals like learning that could occur during first-hand use of a computer-based learning environment. Future research should delve deeper into contextualizing these virtual indicators within classrooms and explore additional contextual factors that could further inform what affects feature interpretation by students. Despite efforts to minimize bias in feature name translations, subjective interpretation may have influenced the final descriptions. Future work could explore more objective methods for feature name translation and investigate the impact of different translations on interpretability. Participants' lack of familiarity with the learning environment may have impacted their perception of feature interpretability. Future research should explore the effects of familiarity on interpretability by conducting studies with participants who have direct experience with the learning environments in question.This study focused primarily on students' perceptions of interpretability, though other constructs such as conceptual understanding, trust, and reliance on machine learning decisions are also important to consider. Future work should explore these issues, including how interpretability moderates the relationship between properties of different features and constructs like trust and reliance.

It is important to acknowledge that interpretability is a multi-dimensional concept and may have two facets: how a feature was created, and why a feature value resulted in a particular ML decision (e.g., why a feature is negatively related to learning performance). In our paper, we combined these two facets of interpretability as a preliminary exploration. While this approach provides valuable insights, it is essential to recognize that a more comprehensive understanding of interpretability would benefit from a multi-faceted investigation that distinguishes between these two aspects.

VI. Conclusion

Our study focused on understanding how well college students can interpret features used in machine learning models for educational outcomes. We found a significant difference in the interpretability of features created by human experts compared to those generated by AutoML methods across multiple datasets. While manual feature engineering is resource-intensive, the disparity in interpretability suggests that human expertise still plays an essential role in crafting understandable features. As machine learning continues to be more deeply integrated into educational platforms, the design and presentation of features demand heightened attention. Of particular note, there is still a need for further research to explore the underlying factors that contribute to these differences in feature interpretability. This will be essential for improving both the accuracy and the user-friendliness of machine learning systems in educational settings. In addition, we recommend methodological advancements to enhance the interpretability of existing AutoML-generated features, and to develop new AutoML feature engineering methods that inherently consider interpretability constraints. We hope our findings will serve as a catalyst for related research, particularly concerning the design and application of features in educational software.

References

[1] M. M.-A.-B. Ahea, M. R. K. Ahea, and I. Rahman. The Value and Effectiveness of Feedback in Improving Students' Learning and Professionalizing Teaching in Higher Education. *Journal of Education and Practice*, 7(16):38–41, 2016. Publisher: IISTE ERIC Number: EJ1105282.

[2] E. Alyahyan and D. Düştegör. Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1):3, Feb. 2020.

[3] N. Bosch. AutoML Feature Engineering for Student Modeling Yields High Accuracy, but Limited Interpretability. *Journal of Educational Data Mining*, 13(2):55–79, Aug. 2021. Number: 2.

[4] S. N. Brohi, T. R. Pillai, S. Kaur, H. Kaur, S. Sukumaran, and D. Asirvatham. Accuracy Comparison of Machine Learning Algorithms for Predictive Analytics in Higher Education. In M. H. Miraz, P. S. Excell, A. Ware, S. Soomro, and M. Ali, editors, *Emerging Technologies in Computing*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pages 254–261, Cham, 2019. Springer International Publishing.

[5] S. Bull. There are Open Learner Models About! *IEEE Transactions on Learning Technologies*, 13(2):425–448, Apr. 2020.

[6] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, Aug. 2019. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[7] C. Conati, K. Porayska-Pomsta, and M. Mavrikis. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling, June 2018. arXiv:1807.00154 [cs].

[8] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning, Mar. 2017. arXiv:1702.08608 [cs, stat].

[9] J. Drozdal, J. Weisz, D. Wang, G. Dass, B. Yao, C. Zhao, M. Muller, L. Ju, and H. Su. Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 297–307, Mar. 2020. arXiv:2001.06509 [cs, stat].

[10] B. Drăgulescu and M. Bucos. Hyperparameter Tuning Using Automated Methods to Improve Models for Predicting Student Success. In A. Lopata, R. Butkienė, D. Gudonienė, and V. Sukackė, editors, *Information and Software Technologies*, Communications in Computer and Information Science, pages 309–320, Cham, 2020. Springer International Publishing.

[11] T. Gervet, K. Koedinger, J. Schneider, and T. Mitchell. When is Deep Learning the Best Approach to Knowledge Tracing? *Journal of Educational Data Mining*, 12(3):31–54, Oct. 2020. Number: 3.

[12] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, Oct. 2018.

[13] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. Number: 1.

[14] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, Aug. 2023.

[15] A. B. Hernández-Lara, A. Perera-Lluna, and E. Serradell-López. Applying learning analytics to students' interaction in business simulation games. The usefulness of learning analytics to know what students really learn. *Computers in Human Behavior*, 92:600–612, Mar. 2019.

[16] M. Hoq, P. Brusilovsky, and B. Akram. Analysis of an Explainable Student Performance Prediction Model in an Introductory Programming Course. In *16th International Conference on Educational Data Mining*, July 2023.

[17] J.-M. John-Mathews. Some critical and ethical perspectives on the empirical turn of AI interpretability. *Technological Forecasting and Social Change*, 174:121209, Jan. 2022.

[18] J. M. Kanter and K. Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, Oct. 2015.

[19] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. 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*, CHI '20, pages 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[20] J. Kay. Scrutable Adaptation: Because We Can and Must. In V. P. Wade, H. Ashman, and B. Smyth, editors, *Adaptive Hypermedia and Adaptive Web-Based Systems*, Lecture Notes in Computer Science, pages 11–19, Berlin, Heidelberg, 2006. Springer.

[21] S. Kelly, S.-A. Kaye, and O. Oviedo-Trespalacios. What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, 77:101925, Feb. 2023.

[22] B. Kim, J. Shah, and F. Doshi-Velez. Mind the Gap: a generative approach to interpretable feature selection and extraction. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2260–2268, Cambridge, MA, USA, 2015. MIT Press.

[23] G. D. Kuh, J. Kinzie, J. Buckley, B. Bridges, and J. C. Hayek. What Matters to Student Success: A Review of the Literature. In *National Symposium on Postsecondary Student Success: Spearheading a Dialog on Student Success*, 2006.

[24] J. Kuzilek, M. Hlosta, and Z. Zdrahal. Open University Learning Analytics dataset. *Scientific Data*, 4:170171, Nov. 2017.

[25] N. A. Levin. Process Mining Combined with Expert Feature Engineering to Predict Efficient Use of Time on High-Stakes Assessments. *Journal of Educational Data Mining*, 13(2):1–15, 2021.

[26] F. Martin and D. Bolliger. Engagement Matters: Student Perceptions on the Importance of Engagement Strategies in the Online Learning Environment. 22:205–222, Mar. 2018.

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

[28] A.-M. Nussberger, L. Luo, L. E. Celis, and M. J. Crockett. Public attitudes value interpretability but prioritize accuracy in Artificial Intelligence. *Nature Communications*, 13(1):5821, Oct. 2022.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay. Scikit-learn: Machine Learning in Python, June 2018. arXiv:1201.0490 [cs].

[30] E. Peer, D. Rothschild, A. Gordon, Z. Evernden, and E. Damer. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4):1643–1662, Aug. 2022.

[31] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach. Manipulating and Measuring Model Interpretability, Aug. 2021. arXiv:1802.07810 [cs].

[32] R. Rapp. Using Collections of Human Language Intuitions to Measure Corpus Representativeness. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2117–2128, Dublin, Ireland, Aug. 2014. Dublin City University and Association for Computational Linguistics.

[33] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura. Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, 21(1):135–146, 2013.

[34] C. Romero and S. Ventura. Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, Nov. 2010.

[35] D. Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146:102551, Feb. 2021.

[36] K. Spoon, J. Beemer, J. C. Whitmer, J. Fan, J. P. Frazee, J. Stronach, A. J. Bohonak, and R. A. Levine. Random Forests for Evaluating Pedagogy and Informing Personalized Learning. *Journal of Educational Data Mining*, 8(2):20–50, Dec. 2016. Number: 2.

[37] T. Susnjak, G. S. Ramaswami, and A. Mathrani. Learning analytics dashboard: a tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, 19(1):12, Feb. 2022.

[38] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos. Implementing AutoML in Educational Data Mining for Prediction Tasks. *Applied Sciences*, 10(1):90, Jan. 2020.

[39] E. Union. EUR-Lex - 32016R0679 - EN - EUR-Lex, 2016.

[40] M. Vahdat, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg. A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator: 10th European Conference on Technology Enhanced Learning, EC-TEL 2015. *Design for Teaching and Learning in a Networked World : 10th European Conference on Technology Enhanced Learning, EC-TEL 2015, Toledo, Spain, September 15–18, 2015 : Proceedings*, pages 352–366, 2015.

[41] F. W. Brown Corpus Manual, 1964.

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

[43] M. Yağcı. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1):11, Mar. 2022.

[44] M. R. Young. The Art and Science of Fostering Engaged Learning. *The Academy of Educational Leadership Journal*, Nov. 2010.

[45] A. Zytek, I. Arnaldo, D. Liu, L. Berti-Equille, and K. Veeramachaneni. The Need for Interpretable Features: Motivation and Taxonomy. *ACM SIGKDD Explorations Newsletter*, 24(1):1–13, 2022.