


Mitigating Cognitive Biases in Predicting Student Dropout: Global and Local Explainability with Explainable Boosting Machine


Rodrigo Costa Camargos

(Universidade Presbiteriana Mackenzie, São Paulo, Brazil)

 <https://orcid.org/0009-0006-7311-342X>, rodrigocamargos@hotmail.com

Ismar Frango Silveira

(Universidade Presbiteriana Mackenzie, São Paulo, Brazil)

 <https://orcid.org/0000-0001-8029-072X>, ismar.silveira@mackenzie.br

Abstract: This study explores the application of Explainable Artificial Intelligence (XAI) techniques to mitigate cognitive biases in predicting student dropout. Focusing on the Explainable Boosting Machine (EBM), we compare its performance and explainability with Logistic Regression and XGBoost models. While EBM and Logistic Regression have inherent explainability, we employ SHAP and Morris Sensitivity Analysis for XGBoost to provide both local and global explanations. Our findings indicate that the inherently interpretable nature of EBM supports clear and actionable decision-making in educational settings. When integrated with additional XAI methods for comparative analysis with models like Logistic Regression and XGBoost, the approach can further enhance the understanding of key factors contributing to student dropout.

Keywords: Learning Analytics, Explainable Artificial Intelligence, Distance Education

Categories: L.3, L.3.5, L.3.6

DOI: 10.3897/jucs.131773

1 Introduction

Predicting student dropout is a critical challenge in education, as it enables the identification of at-risk students and the implementation of preventive interventions. However, the use of complex predictive models, such as XGBoost, can introduce cognitive biases due to their opacity. Cognitive biases are systematic patterns of deviation from norm or rationality in judgment [Tversky, 74], which can affect how users interpret model predictions and make decisions based on them. These biases include overreliance on model outputs, misunderstanding of probabilistic information, and the illusion of explanatory depth, where users believe they understand the model's decision-making process better than they actually do.

Explainable Artificial Intelligence (XAI) aims to address this issue by providing transparency and interpretability to these models [Miller, 19]. This study focuses on the Explainable Boosting Machine (EBM), which inherently offers interpretable predictions [Nori, 19], and compares its performance and explainability with Logistic Regression (LR) and XGBoost [Chen, 16] models. While EBM and LR have built-in explainability, SHAP [Lundberg, 17] and Morris Sensitivity Analysis (MSA) [Morris,

91] are applied to XGBoost to provide both global and local explanations. By leveraging these XAI techniques, we aim to mitigate cognitive biases and enhance the understanding of factors contributing to student dropout. Our findings indicate that the inherently interpretable nature of EBM supports clear and actionable decision-making in educational settings. When integrated with additional XAI methods for comparative analysis with models like LR and XGBoost, the approach can further enhance the understanding of key factors contributing to student dropout.

2 Related Work

The importance of explainability in artificial intelligence has grown significantly, particularly in education, where interpretable predictions enable informed decision-making by educators and administrators. Several studies emphasize the role of explainable models in fostering trust and improving the adoption of AI-driven solutions in this domain. Fiok [Fiok, 22] highlight how explainability enhances stakeholders' confidence in AI systems, while Khosravi [Khosravi, 22] demonstrate the necessity of transparent methods for effective educational interventions.

Beyond the importance of explainability, multiple studies explore different XAI techniques for student dropout prediction. Krüger [Krüger, 23] propose an explainable machine learning approach using SHAP to provide interpretable insights into model decisions. Their findings suggest that incorporating external socioeconomic indicators, such as GDP and the Human Development Index, improves predictive performance. Moreover, their results reinforce SHAP's superiority over LIME [Ribeiro, 16] in terms of stability and interpretability. While their work evaluates multiple machine learning models, our study extends this research by employing EBM and systematically assessing both global and local interpretability in the educational domain.

Similarly, Melo [Melo, 22] explore XAI applications for dropout prediction in a Brazilian technical school. Their study introduces an explainability index to compare LIME, SHAP, and Shapley Values, concluding that SHAP provides more reliable and interpretable explanations for educational stakeholders. Additionally, they emphasize the importance of ethical considerations and fairness when applying AI in education, reinforcing the need for transparent and auditable models. Our research builds upon these insights by using EBM to enhance both interpretability and fairness in student dropout prediction.

In addition to these works, approaches such as DVTXAI [Kamal, 24] demonstrate how advanced deep vision transformers can be coupled with an explainable AI framework (SHAP) to interpret classification tasks in agriculture. Although our current research utilizes tabular data for student dropout prediction, methods like DVTXAI offer valuable insights for extending transformer-based architectures to image datasets. This dual focus on high-performance modeling and interpretability reinforces the importance of adaptable XAI frameworks capable of serving distinct domains, ranging from educational analytics to precision agriculture.

Further studies highlight the broader implications of XAI in educational contexts. Shin [Shin, 21] discusses the impact of explainability and causability on user trust and acceptance of AI systems, underscoring the importance of interpretable models in high-stakes decision-making. Dsilva [Dsilva, 23] validate the effectiveness of EBM in academic risk prediction, demonstrating its ability to generate reliable and interpretable

insights. Likewise, Alamri [Alamri, 21] conduct a systematic review of explainable student performance prediction models, emphasizing the challenge of balancing predictive performance with interpretability.

Addressing the broader debate on the trade-off between interpretability and predictive performance, Kruschel [Kruschel, 25] challenge the conventional assumption that interpretable models inherently sacrifice accuracy. Their study systematically evaluates seven Generalized Additive Models (GAMs), including EBM, and compares them with seven commonly used black-box models across 20 benchmark datasets. The findings reveal that EBM achieves comparable, and sometimes superior, performance to deep neural networks and ensemble methods, demonstrating that interpretable models can be both accurate and transparent. Furthermore, the study highlights the advantages of GAM-based models in maintaining interpretability without relying on post-hoc explanation techniques. These findings reinforce the suitability of EBM in educational predictive modeling, aligning with our research focus on enhancing explainability in student dropout prediction.

Our study builds on these findings by employing EBM for student dropout prediction, integrating SHAP to enhance feature-level explanations. By doing so, we contribute to the growing body of research advocating for explainable models in education. Furthermore, our findings align with previous works [Nori, 19; Krüger, 23], which emphasize the role of explainability in mitigating biases and ensuring fairness in predictive models.

3 Theoretical Foundation

In this section, we provide an overview of the predictive models used in this study: LR, EBM, and XGBoost. We then discuss the XAI techniques employed, specifically SHAP and MSA, and briefly explore cognitive biases, examining their impact on model interpretation and decision-making. Each subsection will delve into the specific characteristics and applications of these models and techniques.

3.1 Logistic Regression (LR)

LR is a widely used statistical technique for modeling and analyzing binary response variables [Berkson, 44]. This method predicts the probability of a dependent binary variable based on one or more independent variables. The central logistic function, given in Equation 1, maps any continuous value to a range between 0 and 1, facilitating its interpretation as a probability. This model adjusts the coefficients β_0 and β_1 through the maximum likelihood method, seeking estimates that maximize the likelihood of the observed data under the proposed model.

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad (1)$$

The likelihood function, expressed in Equation 2, is used to find the coefficients that best fit the model to the data. By transforming the odds ratio $\frac{p(X)}{1-p(X)} = e^{\beta_0 + \beta_1 X}$, we

obtain the logit or logarithm of the odds, simplifying the relationship between the dependent and independent variables to a linear form, $\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X$.

$$l(\beta_0, \beta_1) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \quad (2)$$

The ability of logistic regression to provide probabilistic outputs and its inherent interpretability make it a valuable tool in many scenarios, including classification tasks in the context of student dropout, where decisions often depend on the probability of binary event occurrences.

3.2 Explainable Boosting Machine (EBM)

EBM is an advanced machine learning model known for its high interpretability and accuracy comparable to black-box methods such as Random Forest and Gradient Boosted Trees [Nori, 19]. EBM extends GAMs and is designed to provide intrinsic explainability, making it suitable for classification tasks in educational data science. The mathematical formulation of EBM is described in Equation 3 below:

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i < j} f_{ij}(x_i, x_j) \quad (3)$$

where:

- $g(E[y])$ is the link function that adapts the GAM to different settings, such as regression or classification.
- β_0 is the intercept term.
- $f_j(x_j)$ are individual feature functions that show how each feature x_j independently contributes to the model prediction.
- $f_{ij}(x_i, x_j)$ are interaction terms between pairs of features, capturing the relationship between two features x_i and x_j and how this interaction affects the model prediction.

EBM utilizes modern machine learning techniques, such as bagging and boosting, to learn these feature functions. During training, EBM adopts a round-robin approach to sequentially update the feature functions one feature at a time, using a low learning rate (learning rate: 0.01). This ensures that the order of the features does not influence the final model, minimizes the effects of collinearity, and learns the best function f_j for each feature. Additionally, EBM can automatically detect and include interaction terms between pairs of features, which enhances its accuracy while maintaining interpretability.

3.3 eXtreme Gradient Boosting (XGBoost)

XGBoost is a supervised learning algorithm that improves gradient boosting methods through a series of technical optimizations. Developed by Chen and Guestrin [Chen, 2016], this method combines multiple decision trees to form a final robust and accurate model. It uses the sequential boosting technique, where each new tree is built to correct the errors of the previous trees, as represented by Equation 4.

$$h(x) = \sum_{b=1}^B f_b(x_i), f_b \in \mathcal{F} \quad (4)$$

where B is the number of trees, and f_b represents the functions in the functional space \mathcal{F} .

XGBoost is efficient in processing large volumes of data, supports various loss functions, and includes regularization to prevent overfitting. Its flexible configuration and efficiency in parallel and distributed hardware make it a powerful tool for classification and regression problems.

3.4 SHapley Additive exPlanations (SHAP)

SHAP is a model-agnostic explainer capable of providing both local and global explanations based on game theory. It aims to elucidate a black-box model by calculating the contribution, Shapley values, of each attribute in the prediction [Lundberg, 17], as described in Equation 5. SHAP considers explanations in the form:

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_j z_j' \quad (5)$$

where g is the explainable model, ϕ is the Shapley value, and M is the maximum coalition size. This approach ensures that each feature's contribution to the prediction is fairly and comprehensively represented, aiding in the interpretability of complex models.

3.5 Morris Sensitivity Analysis (MSA)

MSA efficiently identifies influential factors in complex models using a factorial sampling strategy, varying input factors one at a time to estimate their elementary effects [Morris, 91]. It computes two key metrics: the mean absolute effect, indicating overall influence, and the standard deviation, capturing non-linearities and interactions. This method is particularly useful for computationally expensive models, requiring fewer simulations than other global sensitivity analyses. By analyzing these metrics, MSA helps determine the most impactful factors, enabling more focused and efficient model evaluations.

3.6 Cognitive Bias

Cognitive biases are systematic deviations from rational judgment that affect decision-making, particularly in XAI-assisted processes, where users may misinterpret model explanations, leading to errors or overconfidence [Bertrand, 22]. Examples include *confirmation bias*, favoring preconceptions while ignoring alternatives, and *anchoring bias*, over-relying on initial information [Tversky, 74]. Poorly designed explanations can also introduce biases, such as automation bias or the illusion of explanatory depth, where users overestimate their understanding of model decisions [Miller, 19].

Mitigating biases is crucial for reliable XAI systems. Techniques such as counterfactual explanations, highlighting prediction uncertainty, and providing diverse perspectives help reduce their impact [Bertrand, 22; Lundberg, 17]. In high-stakes domains like education and healthcare, tailored explanation frameworks ensure equitable decision-making [Alamri, 21; Fiok, 22].

This study employs EBM to address cognitive biases by offering interpretable global explanations and localized insights, reducing anchoring bias and the illusion of explanatory depth. Complementary methods, such as SHAP for XGBoost and MSA, enhance transparency by quantifying feature contributions and global input sensitivity. Additionally, the TalkToEBM framework translates EBM outputs into natural language, making results accessible to non-technical users.

3.7 Evaluation Metrics for XAI Methods

To validate the findings from different XAI methods used in this study, we adopted a metric-based evaluation inspired by Melo [Melo, 2022]. This evaluation framework consists of 12 predefined metrics designed to systematically assess the explainability of methods such as EBM explanations, SHAP, and MSA. Key metrics include feature importance (M7), weight in decisions (M8), and simplicity of explanations (M10), which were evaluated to highlight the strengths and weaknesses of each method. This systematic comparison provides a robust framework for analyzing the effectiveness of explainability in the context of student dropout prediction. The full set of metrics used in this study is detailed below:

- M1: Understand how inputs are mathematically mapped to outputs
- M2: View the characteristics of parameters
- M3: Visualize interactions with the dataset
- M4: Interactive views
- M5: Understand why one method is better than another
- M6: Understand what can be changed in a model so the output is the desired one
- M7: Show the importance of features
- M8: Show the weight of the features in each decision
- M9: Understand why unobserved events could have occurred
- M10: The presentation of the explanation is simple
- M11: Show probabilities
- M12: Generalization of the AI model

This metric-based approach ensures a comprehensive and objective comparison of different XAI methods, facilitating the identification of suitable techniques for educational contexts.

4 Methodology

In this study, we use EBM as the primary model for student dropout prediction due to its interpretability and strong performance. SHAP values are employed for global and local feature attributions, while MSA evaluates feature impact.

Additionally, we include LIME in our experiments to assess its effectiveness and limitations in providing local explanations, comparing its performance against SHAP.

4.1 Description of Experiments

The experiments began with an Exploratory Data Analysis (EDA) to understand data distributions and relationships among variables, potentially introducing initial cognitive biases. Following this, we trained and evaluated three models: EBM, LR, and XGBoost, assessing their ability to predict student dropout. For XGBoost, SHAP, LIME and MSA were applied to provide global and local explanations, while EBM and LR offered inherent explainability through their coefficients and structures.

Given that the dataset contains only 6655 records, we selected models that balance performance and interpretability effectively. LR and EBM were chosen for their inherent interpretability. More complex models like Random Forest, Support Vector Machines, and Neural Networks were not included, as their performance advantage is often negligible with datasets of this size and complexity, particularly when compared to simpler models like LR [Caruana, 06]. Neural Networks, often used for large-scale and high-dimensional data, have shown potential in educational domains for analyzing unstructured data, such as student-generated content or video interactions. However, their applicability to tabular datasets, like the one used in this study, is limited, and their interpretability remains a challenge [Miller, 19].

Recent studies have explored the use of deep learning models for dropout prediction in higher education, demonstrating competitive performance compared to traditional machine learning approaches [Baranyi, 20; Melo, 22]. However, these models require extensive hyperparameter tuning, significant computational resources, and rely on post-hoc interpretability techniques such as SHAP and LIME to explain their predictions. While recent advances in XAI methods, such as Grad-CAM, Integrated Gradients, and LIME, have improved the interpretability of neural networks, their explanations often lack the simplicity and directness offered by inherently interpretable models like LR and EBM. As highlighted by Baranyi [Baranyi, 20], the lack of transparency in deep learning can hinder its adoption in education, where interpretability is crucial for decision-making. To provide a benchmark for comparison, XGBoost was included as a representative black-box model due to its robustness and effectiveness [Chen, 16], ensuring a comprehensive evaluation that balances performance and interpretability.

Although LIME is a widely used local explanation method, recent studies have highlighted its limitations, particularly in ensuring true locality and stability in tabular data explanations. The method generates explanations by fitting a local surrogate model based on perturbed samples, but these perturbations may not lie within the data manifold, leading to misleading attributions. Additionally, LIME lacks theoretical guarantees regarding the consistency of its explanations. Ghalebikesabi [Ghalebikesabi, 21] argue that LIME's reliance on a global reference distribution can distort local interpretability, making it less reliable in certain scenarios.

Given this context, we will conduct experiments with LIME to empirically assess its limitations and compare its results with those obtained using SHAP. Our objective is to evaluate the stability of the generated explanations, the adequacy of perturbations to the data manifold, and the consistency of feature importance attributions.

To enhance robustness, we split the dataset into 80% for training and 20% for testing and used StratifiedKFold cross-validation with five splits to ensure each fold represented the overall class distribution, thus mitigating class imbalance [Kohavi, 95]. The models were trained and validated, with LR limited to 3000 iterations, EBM utilizing all CPU cores, and XGBoost configured with 300 estimators and depth 5.

4.2 Data and Tools

The dataset used in this study comprised student interaction data from an educational platform at the Universidade de Pernambuco (UPE), including variables such as login frequency, forum activity, and assignment submissions [Ramos, 16]. The analysis was conducted using Python, using libraries such as scikit-learn for model training. The InterpretML package [Nori, 19] was employed for LR and EBM, providing global and local explainer visualizations, and MSA for key feature analysis. Additionally, the SHAP and LIME package was used to generate global and local explainers for the XGBoost model.

Furthermore, the TalkToEBM package was employed to mitigate cognitive biases in interpreting the graphs obtained from the EBM model. TalkToEBM provides a natural language interface that translates EBM graphs into understandable textual descriptions, facilitating clearer and more accessible interpretation of the results [Lengerich, 23; Bordt, 24]. This approach helps ensure that the explanations are consistent and comprehensible to a broader audience, including those with less technical expertise, thereby promoting more balanced and informed decision-making.

Data preprocessing involved handling missing values and converting string attributes to numerical values to ensure model compatibility. EDA was performed to understand data distributions and relationships among variables, guiding the selection of features for the models.

4.3 Measurement of Cognitive Bias

Cognitive biases were measured by evaluating how the explanations provided by different explainers influenced the interpretation of the researchers, who also served as participants in this study. The experiments began with an EDA, which may introduce initial cognitive biases, particularly confirmation bias. Researchers assessed model predictions using various global and local explainers provided by the SHAP, LIME and InterpretML Python libraries, including SHAP and LIME for XGBoost and the explainers in InterpretML for LR and EBM.

To quantitatively assess the comprehensiveness of the explainers, we employed the IE_{XAI} metric, inspired by the framework described in Section 3.7. This metric evaluates the proportion of satisfied explainability criteria out of a predefined set of 12 metrics, such as feature importance (M7), weight in decisions (M8), and simplicity of explanations (M10).

$$IE_{XAI} = \frac{\sum_{i=1}^{12} M_{i_{satisfied}}}{12} \quad (7)$$

As defined in Equation 7, $M_{i_{satisfied}}$ indicates whether a specific metric M_i was fulfilled by the explainer. This framework enabled a systematic comparison of global and local explainers, including SHAP, LIME, MSA, and EBM, providing a robust and quantitative measure of their effectiveness.

While the IEXAI metric enables an objective comparison, it is important to acknowledge that the interpretation of the results still depends on the researchers' subjective analysis. A key limitation of this study is the lack of external validation with domain experts, such as educators and psychologists. Future work could address this limitation by involving these stakeholders to evaluate the practical relevance and interpretability of the model explanations in real-world educational contexts. Such validation would provide additional insights into how effectively the explanations align with domain-specific knowledge and support actionable decision-making.

4.4 Fairness Analysis

Fairness is crucial in educational applications, as algorithmic decisions can reinforce societal biases [Hardt, 2016; Verma, 2018]. This study assessed bias in dropout prediction models to ensure no group is unfairly disadvantaged.

Since the dataset lacked sensitive attributes (e.g., gender, race, or socioeconomic status), we analyzed fairness by: (1) monitoring model performance across courses and subjects; and (2) comparing prediction distributions in academic subpopulations [Friedler, 19]. While not a full demographic fairness analysis, this approach indicates potential biases in curricular groups.

We computed False Positive Rate (FPR) and False Negative Rate (FNR) per subgroup, as disparities in these metrics suggest classification biases [Hardt, 16]. Some models reached 100% accuracy in certain subjects, but small sample sizes (e.g., fewer than 10 instances) inflated error rates. For instance, while “Pedagogia” (Pedagogy) showed near-perfect performance, “Biologia” (Biology) had higher FPR and FNR due to fewer samples.

To address imbalances, possible strategies include: (i) adjusting decision thresholds to balance FPR/FNR; or (ii) re-sampling methods incorporating sensitive attributes when available [Mehrabi, 22]. However, this study focuses on academic engagement variables rather than demographic fairness. Future work will incorporate socioeconomic data and expert validation to align fairness techniques with educational objectives.

5 Results and Discussion

This section presents experimental findings, analyzing key explanatory features and their impact on student dropout rates. We compare the performance and explainability of LR, EBM, and XGBoost, highlighting strengths, limitations, and correlations in the data. Additionally, we discuss the implications for educational data science, focusing on actionable insights to reduce dropout.

The following subsections detail model performance, variable importance, and the effectiveness of XAI techniques in providing transparent and interpretable explanations.

5.1 Exploratory Data Analysis (EDA)

In the EDA, we selected key variables for dropout prediction based on theoretical relevance, data quality, and initial correlations [Ramos, 16]. Variables related to interaction frequency, student engagement, and instructor support were prioritized for a clearer and more actionable model.

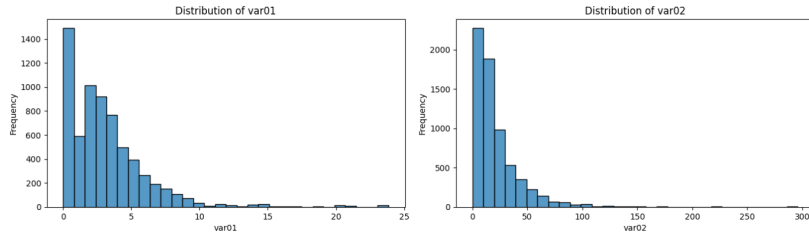


Figure 1: Distribution of var01 and var02

Histograms in Figures 1 and 2 reveal that "var01" (weekly accesses), "var02" (morning accesses), and "var03" (afternoon accesses) follow right-skewed distributions, with most students having low engagement. "Var10" (engagement in activities) also shows low participation, reinforcing the link between limited interaction and higher dropout risk.

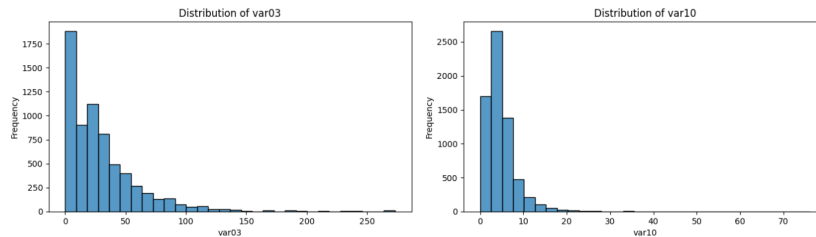


Figure 2: Distribution of var03 and var10

A heatmap of correlations (Figure 3) highlights that "var13," "var10," and "var01" have the strongest negative correlations with dropout (-0.62, -0.48, -0.49), meaning higher engagement reduces dropout likelihood. "Var02" and "var03" also show negative correlations (-0.40, -0.44). "Var17" and "var19" have weaker but still relevant effects (-0.35, -0.32).

High correlations among "var01," "var02," and "var03" suggest possible multicollinearity, as these features represent different engagement dimensions. These findings emphasize the importance of **interaction metrics** in dropout prediction, reinforcing that targeted interventions to boost engagement may help reduce dropout rates.

Table 2 highlights subjects with the highest dropout rates. "Didactics," "Philosophy of Education," "Psychology of Learning," and "Scientific Methodology" each have 39

dropouts, indicating challenges in these areas. "Anthropology," "Educational Assessment," and "Assessment of Learning" follow with 27 dropouts each. These results suggest that certain subjects may be particularly demanding or less engaging, contributing to higher dropout rates. Addressing these challenges could be essential for improving student retention.

Table 3 reveals patterns between student engagement and dropout rates. "Educational Assessment" and "Assessment of Learning" have strong negative correlations with "var01" (-0.591 and -0.534) and "var10" (-0.779 and -0.748), indicating that higher engagement significantly reduces dropout risk. "Didactics" and "Philosophy of Education" also show notable negative correlations with "var01" (-0.592 and -0.504), reinforcing the importance of engagement. In contrast, "Anthropology" has a weaker correlation between "var10" and dropout (-0.112), suggesting that engagement alone may not fully explain dropout in this subject.

Attribute	Description
var01	Average weekly number of student accesses to the virtual learning environment during the semester.
var02	Number of student accesses to the virtual learning environment in the morning shift, per semester.
var03	Number of student accesses to the virtual learning environment in the afternoon shift, per semester.
var10	Number of student accesses to different types of activities provided (webquest, forum, quiz, etc.), per course.
var11	Average weekly number of messages sent by the student within the virtual learning environment, per semester.
var12	Number of student accesses to the forums, per course.
var13	Total number of messages sent by the student within the virtual learning environment, per semester.
var17	Total number of messages received by the student within the virtual learning environment, per semester.
var19	Number of messages from teachers received by the student in the environment, per semester.
var20	Number of messages from peers received by the student within the virtual learning environment, per semester.
var21	Number of messages sent by the student to other peers within the virtual learning environment, per semester.
var26	Total number of resources provided by the teacher (web pages, videos, PDFs, etc.), per course.
var27	Total number of activities provided (webquest, forum, quiz, etc.) by the teacher, per course.
var29	Number of discussion forums provided about the course content, per course.

Table 1: Attributes and Descriptions of the UPE Dataset

These findings highlight that student engagement plays a key role in reducing dropout rates, particularly in subjects related to educational assessment and methodologies. Promoting engagement strategies in high-risk subjects could be an effective approach to improving student retention.

5.2 Model Training and Evaluation

At this stage, the training dataset was balanced using the Random Over Sampler technique from the 'imblearn.over_sampling' Python package [Lemaître, 16] to address class imbalance. This ensured equal representation of both classes, preventing bias toward the majority class. LR, EBM, and XGBoost were then trained using cross-validation to improve robustness and generalizability [Kohavi, 95]. Cross-validation, which partitions the data into multiple training and validation sets, allowed for a thorough evaluation of model performance.

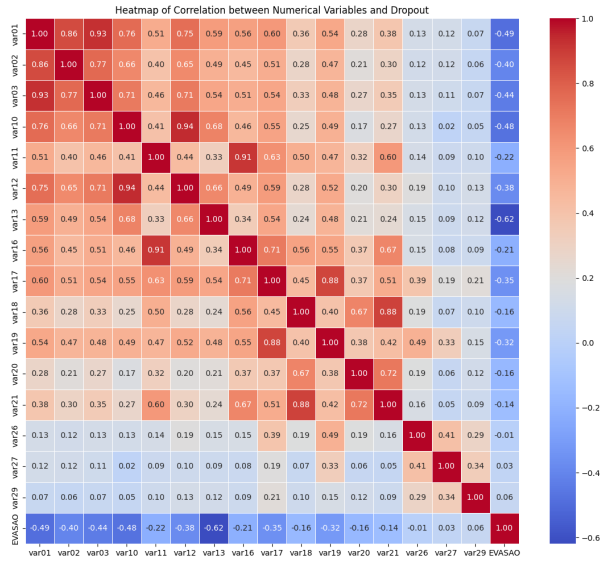


Figure 3: Heatmap of Correlation between Numerical Variables and Dropout

F1 Score and Accuracy were used as performance metrics. The F1 Score balances precision and recall, making it particularly useful for imbalanced datasets, while Accuracy measures the proportion of correctly classified instances. Table 4 presents the results on the test set.

Subject	Dropout
Didactics	39
Philosophy of Education	39
Psychology of Learning	39
Scientific Methodology	39
Anthropology	27
Educational Assessment	27
Assessment of Learning	27

Table 2: Subjects with the Highest Dropout Rates

These results indicate that EBM and XGBoost significantly outperform LR in both F1 Score and Accuracy. The higher F1 Scores and Accuracy values for EBM and XGBoost suggest these models are more effective in correctly identifying both dropout and non-dropout students, thereby providing more reliable predictions.

Subject	Var01 and Dropout	Var10 and Dropout
Anthropology	-0.0374	-0.112
Educational Assessment	-0.591	-0.779
Assessment of Learning	-0.534	-0.748
Didactics	-0.592	-0.372
Philosophy of Education	-0.504	-0.565
Scientific Methodology	-0.389	-0.411
Psychology of Learning	-0.478	-0.519

Table 3: Correlation between Student Engagement and Dropout Rates for High Dropout Subjects

Model	F1 Score	Accuracy
LR	0.868	0.897
EBM	0.993	0.994
XGBoost	0.994	0.995

Table 4: Model Performance

The comparative chart of confusion matrices for LR, EBM, and XGBoost models, shown in Table 5, highlights their performance in predicting student dropouts. LR model shows 911 true negatives and 284 true positives, with some misclassifications (82 false positives and 54 false negatives). The EBM model performs better, with 987 true negatives and 337 true positives, and significantly fewer misclassifications (6 false positives and 1 false negative). XGBoost demonstrates the highest accuracy, with 987 true negatives and 338 true positives, and minimal misclassifications (6 false positives and 0 false negatives). These results indicate that EBM and XGBoost models outperform Logistic Regression in accurately predicting student dropouts.

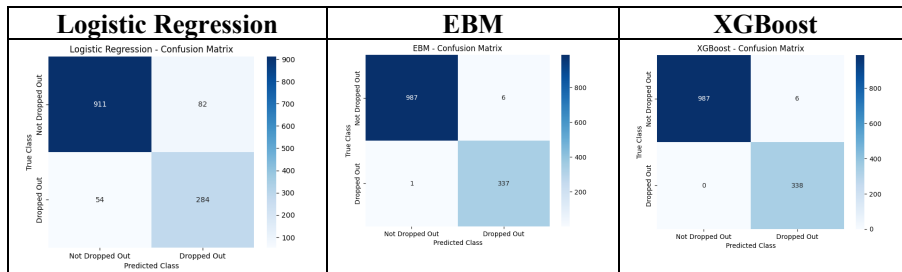


Table 5: Comparative Chart of Confusion Matrices for LR, EBM, and XGBoost Models

5.3 Global Explainers

This section presents the global explainers for LR, EBM, and XGBoost, offering insights into the overall impact of features on dropout prediction. By analyzing coefficients and feature importance, we identify key factors influencing student

retention, guiding targeted interventions. The following subsections provide detailed interpretations for each model.

Figure 4 illustrates the feature importance in LR. "Var11," representing the average weekly number of messages sent in the virtual learning environment, is the most influential feature, strongly reducing dropout risk. Subjects like "Cytology," "Physics Applied to Biology," and "Information Technology Applied to Biology" have significant positive coefficients, indicating a higher dropout likelihood. In contrast, "Pedagogical Practice II" and "Didactics" show negative coefficients, suggesting that engagement in these areas helps reduce dropout. This analysis highlights key factors affecting student retention and informs strategies to improve outcomes.

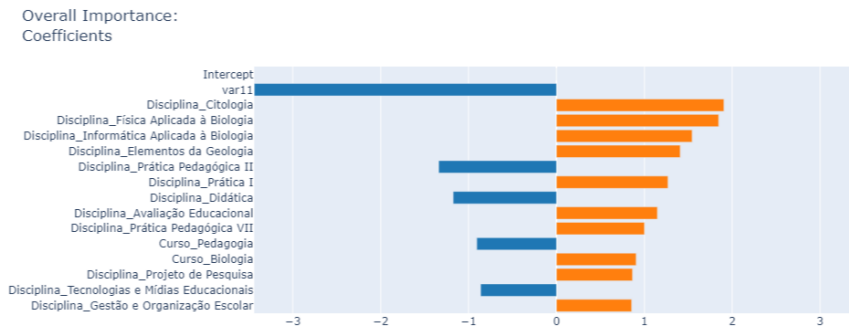


Figure 4: Logistic Regression Global Explainer

Figure 5 highlights the key features influencing dropout in the EBM model. The most influential variables include "var01" (weekly student accesses), "var03" (afternoon accesses), and "var17" (messages received), all showing high mean absolute scores, reinforcing their importance in dropout prediction. Other relevant features include "var19" (messages from teachers), "var20" (messages from peers), and "var11" (messages sent). Combined terms, such as "Period" & "var01" and "Semester" & "var02", illustrate complex interactions affecting student retention.

Comparing the EBM results with the EDA, similar engagement metrics emerge, particularly "var01" (weekly accesses) and "var10" (activity accesses). While EDA emphasizes direct relationships, EBM expands on these findings by identifying new influential variables like "var03" and "var17" and capturing complex feature interactions. However, cognitive bias may occur if researchers rely too heavily on EDA results and overlook the broader insights revealed by EBM, underscoring the need for a comprehensive approach in dropout analysis.

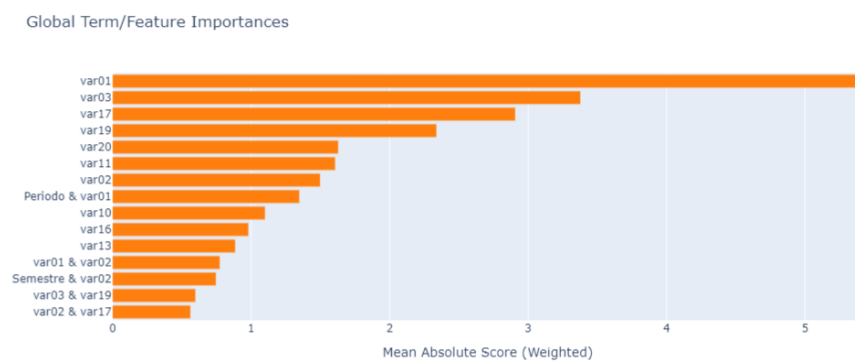


Figure 5: EBM Global Explainer

The strong emphasis on features like "var01" and "var17" in the EBM results reflects the cognitive bias of confirmation, where researchers focus on data supporting pre-existing beliefs about student engagement as a primary factor in dropout. To mitigate this bias, it is essential to consider broader insights from EBM, including interactions between periods and semesters. These findings underscore the importance of engagement metrics and provide actionable insights for developing targeted interventions to enhance student retention.

Traditional feature importance methods in XGBoost, such as Weight, Gain, and Cover, often yield inconsistent results due to differing evaluation perspectives. SHAP and MSA address this by providing more robust insights: SHAP quantifies feature contributions for global and local interpretability, while MSA assesses sensitivity to input variations. Combining these methods enhances reliability, overcoming traditional metric limitations and improving decision-making.

The SHAP summary plot in Figure 6 provides a detailed view of how each feature influences the XGBoost model's predictions. Unlike traditional feature importance charts, this plot displays a density scatter plot of SHAP values, where features are sorted by their overall impact, measured by the sum of SHAP value magnitudes. The feature "var01", representing the average weekly number of student accesses, is the most influential, with a wide range of SHAP values. Higher values of "var01" consistently decrease dropout likelihood, highlighting the critical role of regular engagement.

Other impactful features include "var03" (afternoon accesses) and "var13" (total messages sent by students), both showing strong variations in SHAP values that reflect their significant effects on predictions. Additional key features, such as "Period" (Período), "var11" (average weekly messages sent), and "var02" (morning accesses), further emphasize the importance of engagement and interaction patterns. For example, higher values of "var11" are typically associated with lower dropout predictions. Features like "Curso_Biologia" (Biology course) and "Semester" (Semestre) also play significant roles, with their impacts varying depending on specific feature values.

The color gradient in the plot, ranging from blue (low feature values) to red (high feature values), illustrates nuanced feature impacts across individual samples. Compared to traditional feature importance metrics (gain, weight, cover), the SHAP summary plot offers a more comprehensive view by capturing feature interactions and their effects. This mitigates cognitive biases, ensuring a balanced and objective

interpretation of the model’s behavior and providing researchers with deeper insights into the predictors of student dropouts.

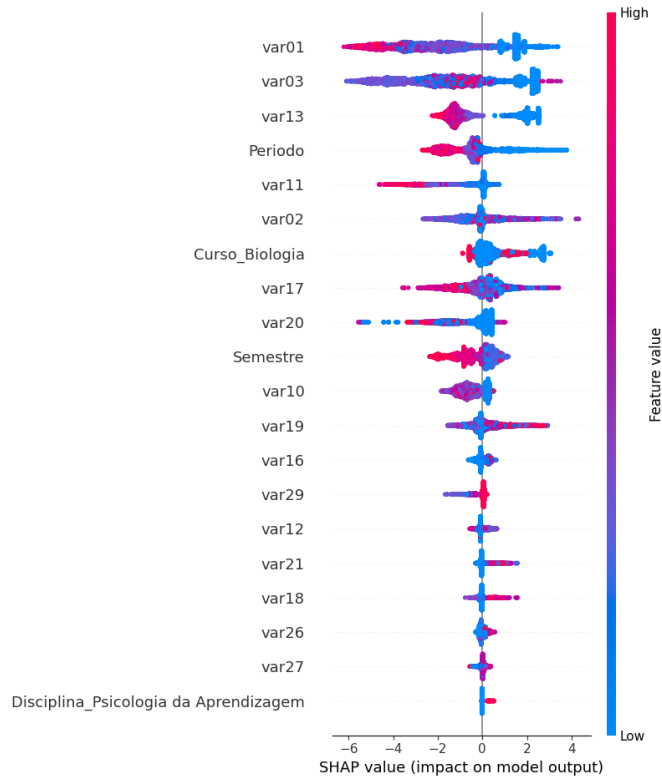


Figure 6: XGBoost - SHAP Global Explainer

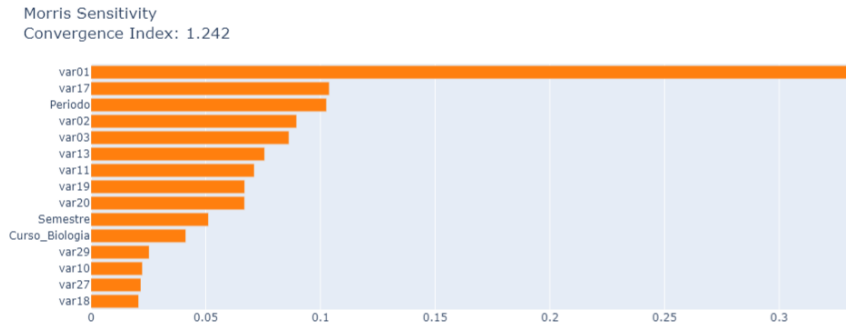


Figure 7: XGBoost - Morris Sensitivity Global Explainer

Figure 7 presents the MSA results for the XGBoost model, showing a convergence index of 1.242, indicating stable results. "Var01" is the most influential feature, reinforcing its importance in dropout prediction, as also seen in the SHAP analysis. Other highly sensitive features include "var17" (messages received), "Period"

(Período), and "var02" (morning accesses), with small changes in their values significantly impacting predictions.

Comparing MSA with SHAP, both methods identify "var01", "var17", and "var03" as key features, though Morris also highlights "Period" and "var02" with slightly different rankings. While SHAP provides detailed contributions at global and local levels, Morris quantifies overall model sensitivity, offering a complementary perspective on feature importance.

5.4 Local Explainers

This section examines local explainers for the three models, which clarify individual predictions. Methods like SHAP, LIME and EBM's built-in explanations show how each feature influences specific outcomes, helping to detect biases, inconsistencies, and unexpected model behavior.

Analyzing local explanations can reveal biases not apparent in global analyses, ensuring fair and transparent predictions. This approach enhances interpretability, mitigates cognitive biases, and supports the development of reliable decision-making frameworks.

The local explainer for the LR model in Figure 8 shows that the predicted outcome for instance eleven is 1 (dropped out), matching the actual value. "Semester" (Semestre) is the most influential feature, strongly increasing dropout risk, while "Period" (Período) and "var27" reduce it. Other features, such as "var29" and "Curso_Biologia," positively contribute to dropout likelihood, whereas engagement metrics like "var01," "var03," and "var11" have minimal impact in this instance.

The local explainer for the LR model in Figure 9 highlights the key features influencing the prediction for instance eighteen in the test set, where the model predicts a value of 0 (not dropped out). "Semester" (Semestre) has a strong positive impact, reducing the likelihood of dropout, as indicated by the prominent orange bar. Conversely, "var11" and other features such as "Period" (Período) and "var13" exert negative impacts, suggesting nuanced interactions that further influence the prediction.

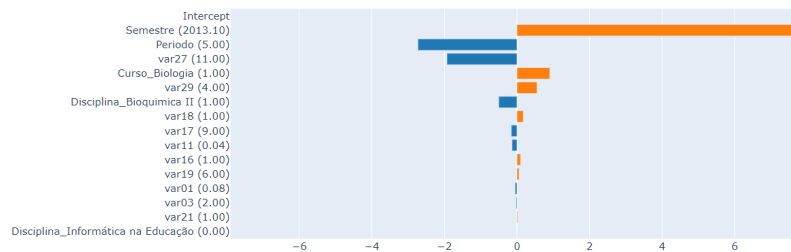


Figure 8: LR - Local Explainer - Instance Eleven

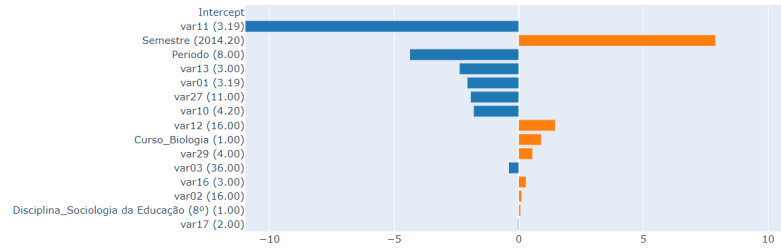


Figure 9: LR - Local Explainer - Instance Eighteen

Local Explanation (Actual Class: 1 | Predicted Class: 1
Pr(y = 1): 1.000)

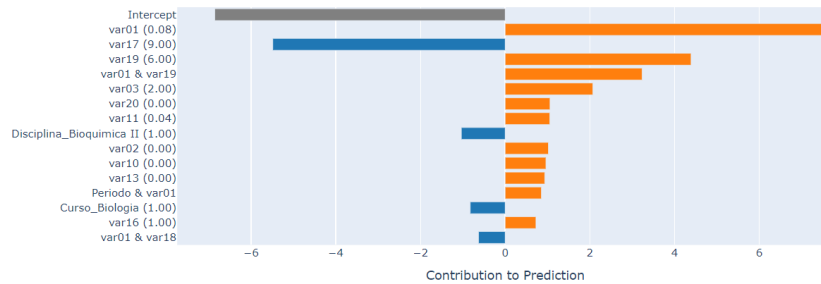


Figure 10: EBM - Local Explainer - Instance Eleven

The local explainer for the EBM model in Figure 10 shows the contributions of each feature to the prediction for instance eleven in the test set, where the model predicts a value of 1 (dropped out) with a probability of 1, matching the actual outcome. "Var01" has the strongest positive impact, significantly increasing the likelihood of dropout, followed by "var19". Conversely, "var17" has the strongest negative impact, reducing the likelihood of dropout. Additionally, the interaction between "var01 & var19" reinforces the prediction. These insights highlight the importance of "var01" and "var17" and their interactions in determining this outcome.

Local Explanation (Actual Class: 0 | Predicted Class: 0
Pr(y = 0): 1.000)

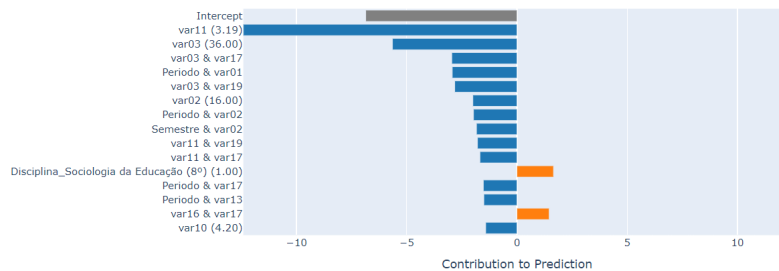


Figure 11: EBM - Local Explainer - Instance Eighteen

The local explainer for the EBM model, shown in Figure 11, provides insights into the contribution of each feature to the prediction for instance eighteen of the test set. The predicted class is 0 (not dropped out) with a prediction probability of 0, matching the

actual class of 0. Key features influencing the prediction include “var11” and “var03”, both of which have significant negative impacts on the prediction, as indicated by the blue bars. This suggests that higher values in these features decrease the likelihood of dropout for this student. Additionally, interactions between “var03 & var17” and “Period & var01” also contribute negatively, reinforcing the prediction that the student will not drop out.

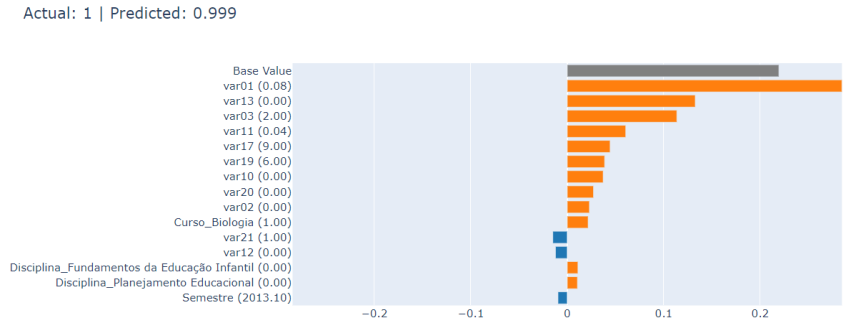


Figure 12: XGBoost - SHAP Local Explainer - Instance Eleven

The SHAP analysis for the XGBoost model provides insights into the feature contributions for individual predictions in the test set. For instance eleven (Figure 12), the predicted class is 1 (dropped out) with a probability of 0.999, matching the actual class. Key features such as “var01,” “var13,” “var03,” and “var11” show significant positive impacts, indicating that higher values in these features increase the likelihood of dropout. This highlights how high engagement metrics strongly influence prediction.

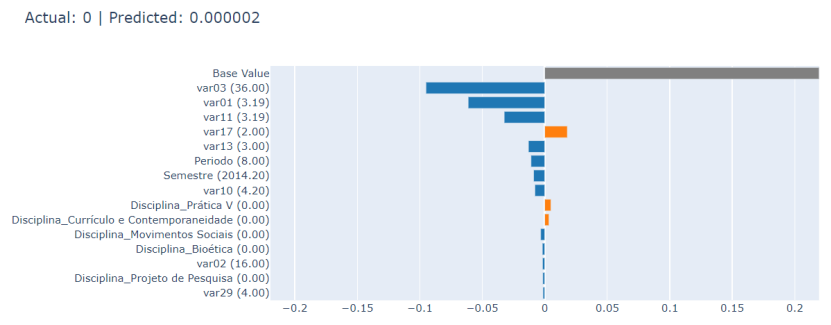


Figure 13: XGBoost - SHAP Local Explainer - Instance Eighteen

For instance eighteen (Figure 13), the predicted class is 0 (not dropped out) with a probability of 0.000002, aligning with the actual class. Features like “var03,” “var01,” and “var11” have significant negative impacts, suggesting that higher values reduce the likelihood of dropout. In contrast, “var17” shows a minimal positive contribution. These results emphasize the nuanced role of individual feature values in shaping predictions.

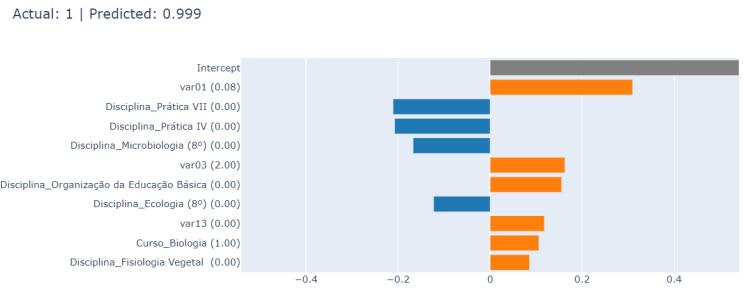


Figure 14: XGBoost - LIME First Execution - Instance Eleven

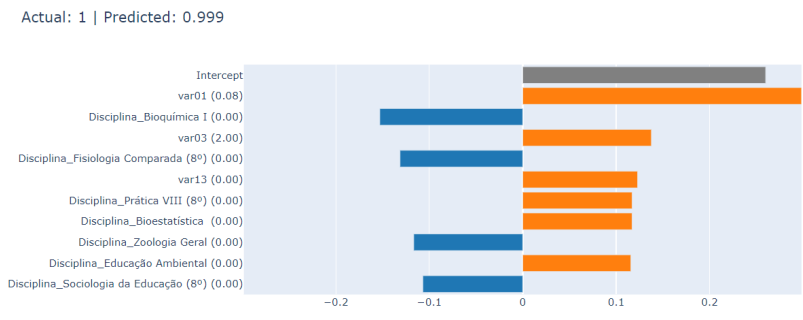


Figure 15: XGBoost - LIME Second Execution - Instance Eleven

Across two separate LIME executions (Figure 14 and Figure 15) for instance eleven, we observed significant inconsistencies in feature attributions, highlighting its instability. While both runs identified "var01" and "var03" as important, the remaining features varied considerably, with different disciplines appearing as either positive or negative contributors in each execution. This instability contrasts sharply with SHAP, which consistently identified "var01," "var13," "var03," and "var11" as key dropout predictors, maintaining a stable ranking of feature importance. The fluctuating attributions in LIME demonstrate its sensitivity to perturbations in local surrogate models, leading to unreliable explanations. In contrast, SHAP provides a robust, theoretically grounded approach, ensuring consistent and interpretable feature contributions. These findings reinforce that SHAP is a more reliable method for explainability in student dropout prediction.

LIME's local explanations were excluded from the IEXAI metric and Table 6 due to their high variability across multiple executions, leading to inconsistent feature attributions. This instability undermines their reliability in comparative assessments, making SHAP more suitable for evaluating explainability.

5.5 Mitigating Cognitive Biases with TalkToEBM

Incorporating TalkToEBM into our study offers significant benefits for mitigating cognitive biases. This open-source package provides a natural language interface to EBMs, allowing us to convert EBM graphs into text comprehensible by Large Language Models (LLMs) [Bordt, 24]. By generating clear, detailed textual

descriptions and summaries of individual graphs or entire models, TalkToEBM enhances the interpretability of EBM’s outputs.

This approach not only facilitates a more transparent and accessible understanding of the model's behavior but also supports a broader range of users, including those with limited technical expertise. Consequently, leveraging TalkToEBM can help ensure more balanced, unbiased interpretations and decisions by presenting diverse perspectives on the model's outputs.

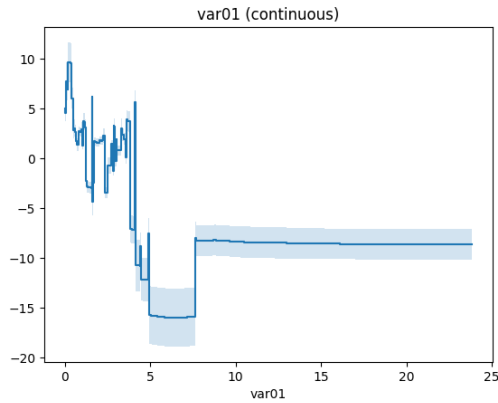


Figure 16: EBM - Graph for Var1 Attribute

Metric	EBM (Global)	EBM (Local)	SHAP (Global)	SHAP (Local)	Morris (Global)	TalkToEBM (Global)
M1	X	X	X	X		
M2	X	X				
M3	X	X				
M4			X	X		
M5	X	X	X	X	X	
M6	X	X	X	X		
M7	X	X	X	X	X	X
M8		X		X		
M9						
M10	X	X			X	X
M11	X	X		X		X
M12	X		X		X	
IEXAI	0.75	0.75	0.50	0.58	0.33	0.25

Table 6: Comparison of Explainability Metrics Across XAI Methods (EBM, SHAP, Morris, and TalkToEBM)

Using the TalkToEBM package, we generated a graph for "var01" (Figure 16) and extracted its interpretation with GPT-4, shown in Table 7. The response explains the non-linear relationship between "var01" and dropout probability, highlighting that both

very low and very high engagement can be detrimental to retention, with specific trends and uncertainties for different levels of engagement.

Table 6 summarizes the comparison of global and local explainability methods. EBM excels in showing feature importance (M7), providing interpretable weights (M8), and offering probabilities (M11). SHAP enhances EBM with detailed post-hoc explanations for local predictions but has limitations in interactivity (M4) and simplicity (M10) for non-technical users. MSA is effective for global evaluations (M1, M2, M7) but lacks support for local explanations.

TalkToEBM complements EBM by converting explanations into natural language, improving accessibility (M10) and interactivity (M4) for non-technical stakeholders. Together, EBM and TalkToEBM create a balanced framework that enhances both interpretability and communication.

6 Conclusions and Future Work

This study aimed to mitigate cognitive biases in predicting student dropout using XAI techniques. We focused on EBM, comparing its performance and interpretability with LR and XGBoost. By integrating SHAP and MSA for XGBoost, we provided both global and local explanations, enhancing understanding of predictive factors and supporting more accurate, fair decision-making in education.

EBM and LR were chosen for their inherent interpretability, essential for actionable insights, while XGBoost served as a high-performing black-box benchmark. Unlike Ramos's study, which used LR but identified different dropout predictors with the same dataset, our findings showed that EBM achieved similar accuracy to XGBoost while offering superior explainability. This transparency enables stakeholders to understand predictions, fostering trust and practical implementation.

Our results identified key dropout predictors, particularly engagement metrics like weekly student accesses, which can guide targeted interventions. Differences in identified predictors compared to previous studies highlight the variability in model selection and feature importance evaluation. Leveraging interpretable models like EBM and LR allows educators to better address dropout risks, while XGBoost ensures robust predictive capabilities.

Despite its contributions, this study has limitations. The findings may not generalize beyond this dataset, and while EBM enhances interpretability, its complexity may still challenge non-technical stakeholders. Additionally, researchers' biases in selecting models and explainers could influence results, underscoring the need for diverse perspectives and further validation. Future research should apply these models to broader educational contexts, incorporate additional behavioral and academic variables, and refine interpretability techniques to improve student retention strategies.

GPT-4 Generated Text
<p><i>The graph of feature var01 was simplified by 0.4%. The GAM graph for the feature `var01`, representing average weekly student accesses to an educational platform, exhibits a complex non-linear relationship with the likelihood of student dropout. Initially, the effect of `var01` on the response variable is positive up to about 0.4, with a notable peak between 0.06 to 0.17, suggesting moderate engagement correlates with reduced dropout probability. However, from around 0.4 to 2.9, the effect becomes unstable, showing rapid fluctuations between positive and negative values, indicating varying impacts of engagement levels within this range. Surprisingly, a sharp and sustained decline in the effect occurs from about 4.17 onwards, reaching extremely negative values by 4.9 and continuing negatively as `var01` increases further. This suggests that very high levels of platform engagement are strongly associated with an increased likelihood of dropout, potentially due to burnout or ineffective cramming strategies. The confidence intervals widen in areas of sharp changes, reflecting greater uncertainty in these regions. This pattern highlights the importance of optimal engagement levels, suggesting that both very low and very high engagement could be detrimental to student retention.</i></p>

Table 7: GPT-4 Analysis of var01 graph using TalkToEBM

We acknowledge the importance of external validation. Our study already employs the IE_{XAI} metric for structured evaluation of explainability. However, we recognize that further validation should involve real-world users. A possible approach is to conduct an empirical study where educators and administrators analyze model predictions with and without explanations, assessing their clarity and usefulness for decision-making. Their feedback can be collected through structured questionnaires, allowing us to measure whether the explanations effectively improve understanding and provide actionable insights. This validation step will ensure that explanations generated by the model are truly interpretable for non-technical users.

References

- [Alamri, 21] Alamri, R., Alharbi, B.: Explainable Student Performance Prediction Models: A Systematic Review, *IEEE Access*, 9, 33132–33143, 2021. <https://doi.org/10.1109/ACCESS.2021.3061368>.
- [Baranyi, 20] Baranyi, M., Nagy, M., Molontay, R.: Interpretable Deep Learning for University Dropout Prediction, In Proc. SIGITE 2020 – 21st Annual Conference on Information Technology Education, 2020. <https://doi.org/10.1145/3368308.3415382>.
- [Berkson, 44] Berkson, J.: Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, 39(227), 357-365, 1944.
- [Bertrand, 22] Bertrand, A., Belloum, R., Eagan, J., & Maxwell, W.: How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, Aug 2022, Oxford, United Kingdom. <https://doi.org/10.1145/3514094.3534164>
- [Bordt, 24] Bordt, S., Lengerich, B., Nori, H., Caruana, R.: Data Science with LLMs and Interpretable Models, arXiv preprint arXiv:2402.14474, 2024. <http://arxiv.org/abs/2402.14474>.
- [Caruana, 06] Caruana, R., Niculescu-Mizil, A.: An Empirical Comparison of Supervised Learning Algorithms, In Proc. 23rd Int. Conf. on Machine Learning (ICML '06), 161–168, 2006.

- [Dsilva, 23] Dsilva, V., Schleiss, J., Stober, S.: Trustworthy Academic Risk Prediction with Explainable Boosting Machines, In Proc. Artificial Intelligence in Education, Lecture Notes in Computer Science, Springer Nature Switzerland, 463–475, 2023. https://doi.org/10.1007/978-3-031-36272-9_38.
- [Fiok, 22] Fiok, K., Farahani, F.V., Karwowski, W., Ahram, T.: Explainable Artificial Intelligence for Education and Training, Journal of Defense Modeling & Simulation, 19(2), 133–144, 2022. <https://doi.org/10.1177/15485129211028651>.
- [Friedler, 19] Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In: FAT '19: Conference on Fairness, Accountability, and Transparency, January 2019, Atlanta, GA, USA. ACM, New York, NY, USA, pp. 329–338. <https://doi.org/10.1145/3287560.3287589>
- [Ghalebikesabi, 21] Ghalebikesabi, S., Ter-Minassian, L., Diaz-Ordaz, K., Holmes, C.: On Locality of Local Explanation Models. In: Advances in Neural Information Processing Systems 34 (NeurIPS 2021), December 2021, pp. 29140–29151.
- [Hardt, 16] Hardt, M., Price, E., Srebro, N.: Equality of Opportunity in Supervised Learning. In: Advances in Neural Information Processing Systems 29 (NIPS 2016), December 2016, pp. 3315–3323.
- [Kamal, 24] Kamal, S., Sharma, P., Gupta, P.K., Siddiqui, M.K., Singh, A., Dutt, A.: DVTXAI: A Novel Deep Vision Transformer with an Explainable AI-based Framework and its Application in Agriculture, The Journal of Supercomputing, 81(280), 2025. <https://doi.org/10.1007/s11227-024-06494-y>.
- [Khosravi, 22] Khosravi, H., et al.: Explainable Artificial Intelligence in Education, Computers and Education: Artificial Intelligence, 3, 100074, 2022. <https://doi.org/10.1016/j.caeai.2022.100074>.
- [Kohavi, 95] Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, In Proc. 14th Int. Joint Conf. on Artificial Intelligence, 1137–1143, 1995.
- [Krüger, 23] Krüger, J.G.C., Britto, A. de S., Barddal, J.P.: An Explainable Machine Learning Approach for Student Dropout Prediction, Expert Systems with Applications, 2023. <https://doi.org/10.1016/j.eswa.2023.120933>.
- [Kruschel, 25] Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M., & Zschech, P. (2025). Challenging the Performance-Interpretability Trade-Off: An Evaluation of Interpretable Machine Learning Models. Business & Information Systems Engineering. <https://doi.org/10.1007/s12599-024-00922-2>
- [Lemaître, 16] Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research, 18(1), 559–563, 2016.
- [Lengerich, 23] Lengerich, B.J., Bordt, S., Nori, H., Nunnally, M.E., Aphinyanaphongs, Y., Kellis, M., Caruana, R.: LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs, arXiv preprint arXiv:2308.01157, 2023. <http://arxiv.org/abs/2308.01157>.
- [Lundberg, 17] Lundberg, S.M., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, In Proc. Advances in Neural Information Processing Systems, 2017.
- [Mehrabi, 22] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning, ACM Computing Surveys, 54(6), 1–35, 2022. <https://doi.org/10.1145/3457607>.

- [Melo, 22] Melo, E., Silva, I., Costa, D.G., Viegas, C.M.D., Barros, T.M.: On the Use of eXplainable Artificial Intelligence to Evaluate School Dropout, *Education Sciences*, 12(12), 845, 2022. <https://doi.org/10.3390/educsci12120845>.
- [Miller, 19] Miller, T.: Explanation in Artificial Intelligence: Insights from the Social Sciences, *Artificial Intelligence*, 267, 1–38, 2019. <https://doi.org/10.1016/j.artint.2018.07.007>.
- [Nori, 19] Nori, H., Jenkins, S., Koch, P., Caruana, R.: InterpretML: A Unified Framework for Machine Learning Interpretability, *arXiv preprint arXiv:1909.09223*, 2019. <https://doi.org/10.48550/arXiv.1909.09223>.
- [Ramos, 16] Ramos, J.L.C.: Uma abordagem preditiva da evasão na educação a distância a partir dos construtos da distância transacional, PhD Thesis, Universidade Federal de Pernambuco, 2016. <https://repositorio.ufpe.br/handle/123456789/21052>.
- [Ribeiro, 16] Ribeiro, M. T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '16)*, 1135–1144, 2016. <https://doi.org/10.1145/2939672.2939778>.
- [Shin, 21] Shin, D.: The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI, *International Journal of Human-Computer Studies*, 146, 102551, 2021. <https://doi.org/10.1016/j.ijhcs.2020.102551>.
- [Tversky, 74] Tversky, A., Kahneman, D.: Judgment under Uncertainty: Heuristics and Biases, *Science*, 185(4157), 1124–1131, 1974.
- [Verma, 18] Verma, S., Rubin, J.: Fairness Definitions Explained, In *Proc. Int. Workshop on Software Fairness, ICSE '18: 40th Int. Conf. on Software Engineering*, Gothenburg, Sweden, 2018, 1–7. <https://doi.org/10.1145/3194770.3194776>.