CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2024.327

A multivariate analysis of the early dropout using classical machine learning and local interpretable model-agnostic explanations

Hai Thanh Nguyen[1*], Phuong Le[2], Tuyen Thanh Thi Nguyen[3], and Anh Kim Su[3]
[1]*College of Information and Communication Technology, Can Tho University, Viet Nam*
[2]*The School of Industrial Biology, France*
[3]*College of Rural Development, Can Tho University, Viet Nam*
**Corresponding author (nthai.cit@ctu.edu.vn)*

Article info.

Received 16 Jun 2024
Revised 4 Sep 2024
Accepted 20 Sep 2024

Keywords

Dropout Prediction, Machine learning, Explanation

ABSTRACT

Student dropout rates can have a significant negative impact on both the development of educational institutions and the personal growth of students. Consequently, many institutions are focused on identifying key factors that contribute to dropout and implementing strategies to mitigate them. This study aims to predict student dropout rates using classical machine learning algorithms while analyzing the key factors influencing these outcomes in higher education. The dataset includes demographic, socioeconomic, and academic information from various sources. Additionally, the study leverages the Local Interpretable Model-Agnostic Explanations (LIME) model to provide insights into the predictions, offering a clearer understanding of the factors driving dropout decisions. This knowledge is crucial for identifying influential factors and, more importantly, enhancing early intervention strategies and policies in educational settings, ultimately reducing dropout rates.

1. INTRODUCTION

Risk of student dropout is a huge issue in higher education in many countries, affecting not just individual prospects, but universities and even the wider socio-economic factors. Higher education institutions face a significant challenge to stem the tide of these educational dropouts. In Spain, the overall dropout rate was 33.9% in 2013 with a total of 35% for public universities and 27.5% for private ones (Oreopoulos & Ford, 2019). Due to significant investment in education, the rate of leaving is 9% in Finland (Vaarma & Li, 2024) and 12% in Korea (Song et al., 2023). The study aimed to identify predictors of student dropout and develop accurate prediction models using advanced machine learning techniques. It leveraged processed data from a published study by (Realinho et al., 2022).

The dataset contained 4424 records with 35 attributes (Realinho et al., 2022). It included comprehensive information collected from many sources, covered student demographics, economics, academic background, and performance. The dataset was from the university's learning management system, instructional support system, and external sources, such as government economic databases and population records. It also included information about students enrolled in university courses in various disciplines, including agriculture, design, education, health, journalism, administration, social services, and technology. In addition to machine learning algorithms Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) and Categorical Boosting (CatBoost), this research applies Local Interpretable Model-Agnostic Explanations (LIME) to interpret the models' predictions. This model provides a transparent way for predicting the behaviour of

dropouts and reveals insights into the dropout mechanisms.

The remaining structure of this article is organized as follows. Section 2 *out*lines the implementation materials and methods used in the study. Section 3 presents the results and discussion. Finally, section 4 concludes by summarizing the key insights and some limitations of this study.

2. MATERIALS AND METHOD

RF, XGBoost. LightGBM and CatBoost are the four machine learning algorithms we use in this ensemble model. The algorithms were trained and evaluated with metrics like accuracy, precision, recall and F1-score to ensure a robust performance. In addition, we used the LIME algorithm to interpret how these models arrive at their decisions. LIME helps to interpret complex models by approximating them locally with simpler, interpretable models. Doing this also helped us to understand how these individual features affect the likelihood of students dropping out. The integration of those modeling techniques and interpretative methods offered a holistic approach to identifying at-risk students, as well as understanding factors contributing to dropout rates.

2.1. Data description

Table 1. Data description

Group	Attribute
Demographic data	Marital status Nationality Displaced Gender Age at enrollment International
Socioeconomic data	Mother's qualification Father's qualification Mother's occupation Father's occupation Educational special needs Debtor Tuition fees up to date Scholarship holder
Macroeconomic data	Unemployment rate Inflation rate GDP
Academic data at enrollment	Application mode Application order Course Daytime/evening attendance Previous qualification
Academic data at the end of 1st semester	Curricular units 1st sem. (credited) Curricular units 1st sem. (enrolled) Curricular units 1st sem. (evaluations) Curricular units 1st sem. (approved) Curricular units 1st sem. (grade) Curricular units 1st sem. (without evaluations)
Academic data at the end of 2nd semester	Curricular units 2nd sem. (credited) Curricular units 2nd sem. (enrolled) Curricular units 2nd sem. (evaluations) Curricular units 2nd sem. (approved) Curricular units 2nd sem. (grade) Curricular units 2nd sem. (without evaluations)

Table 1 shows the dataset (Realinho et al., 2021) of grouped and processed attributes. The dataset includes three target groups as follows: graduate accounts for 49.9%, enrolled accounts for 17.9%, and dropout accounts for 32.1%. This dataset also includes information on personal characteristics (marital status, nationality, gender, and age at admission), socioeconomic status (parents' education level, parents' occupation, student debt status, scholarships), country's macroeconomic data (unemployment rate, inflation rate, and GDP), academic information at the time of admission (application form, courses, and previous education), academic information at the end of the first semester (number of course credits, grades), and information at the end of the second semester (similar to the first semester). As noticed, an imbalance in the target group presence probably affects the models.

This is also a common challenge in many practical problems and often occurs in dropout research. However, it is still valuable for analyzing and interpreting models when the data are unbalanced, primarily when pointing out the influencing factors that lead to dropout. This study is only interested in the factors that influence dropout. On the other hand, we used appropriate evaluation metrics (such as precision, recall, and F1-score) to ensure that the model performance is evaluated relatively despite the imbalance.

The correlation heatmap in Figure 1 illustrates the degree of correlation between features in the dataset. Positive correlations are in red, with values ranging from 0 to 1. The correlation's strength gradually increases with the color's intensity of the color and the value. On the contrary, negative correlations are in blue, ranging from 0 to -0.4; increasing the color's intensity goes with the value's decrease. White-colored cells indicate a very weak correlation or no correlation relationship.

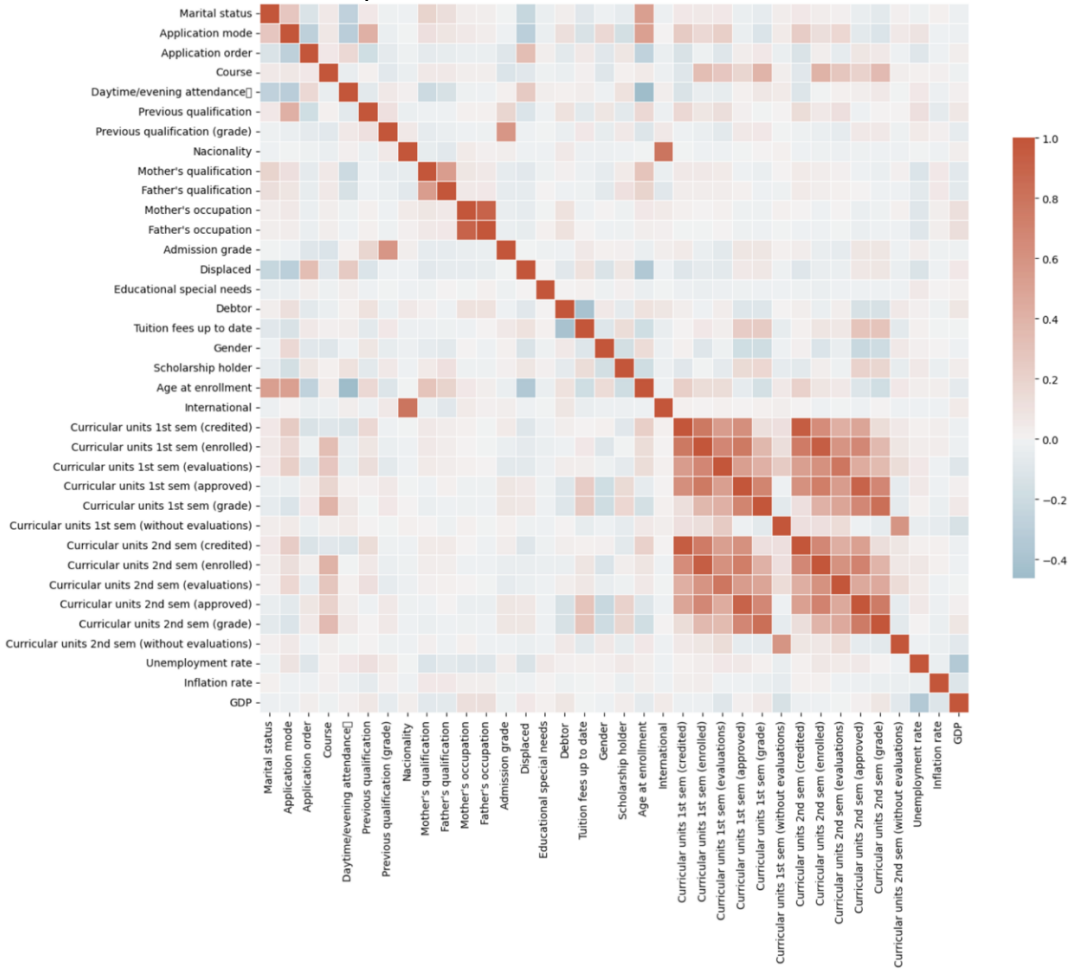

Figure 1. Correlation matrix of the considered features

2.2. Models for the prediction

We use RF, XGBoost, LightGBM, and CatBoost ensemble models instead of single models' decision tree, naive Bayes classifier, SVM, Logistic Regression, etc. for the following reasons: Firstly, increase accuracy because ensemble models are the combination of many single models together, so it can reduce errors and improve accuracy, especially in the case of complex or noisy data. Secondly, reduce variance and bias: for example, Decision Tree models, if not well controlled, often tend to overfit, or Logistic Regression models can have a high bias if the assumptions are unsuitable for the data. Besides, ensemble models do well in this issue. Thirdly, it increases stability and generalization ability. Fourthly, have high flexibility and customization. To ensure objective and accurate evaluation, the data is divided into two sets: 80% for the training set and 20% for the test set. The training set is used to train the model, while the test set evaluates the model's performance on unseen data. We used 10-fold cross-validation for all models in this study.

2.2.1. RF

RF (Rigatti, 2017) An ensemble-based machine learning model uses a series of decision trees to form an aggregation that predicted more accurately and stabler. Decision trees act as base classifier with bagging. Every decision tree is developed by a sub-sample of the training data. The final outcome is selected by taking the average or majority votes of the trees from the forest. The Random Forest model is well-known because it helped handle large datasets, is capable of multi-classification and shows the degree of importance for each feature so that many applications like this model.

2.2.2. XGBoost

XGBoost (Belyadi & Haghighat, 2021) is a robust machine-learning algorithm that uses weak learners as its base classifier and is developed based on the boosting method. XGBoost optimizes the learning process using gradient-boosting techniques on decision trees. The algorithm is faster and handles large and complex data sets. XGBoost enables the automatic handling of missing values, computes higher efficiency, and resists over-fitting due to the integration of regularization methods.

2.2.3. LightGBM

LightGBM (Ke et al., 2017) is a powerful and advanced machine learning algorithm designed to optimize speed and performance when processing large data sets. LightGBM has advanced features, such as fast training speed, scalability, and high performance. LightGBM's strength is its ability to handle heterogeneous data and missing values without complex data preprocessing. LightGBM uses weak learners and decision tree boosting, like XGBoost, with additional speed and memory usage improvements.

2.2.4. CatBoost

CatBoost (Prokhorenkova et al., 2018) is a machine learning algorithm belonging to the gradient boosting group developed by Yandex, a sizeable Russian technology company. CatBoost stands for "Categorical Boosting," which reflects its exceptional ability to handle categorical variables. Like XGBoost and LightGBM, CatBoost uses weak learners and applies the Boosting method.

2.3. Local Interpretable Model-Agnostic Explanations

LIME (Singh & Guestrin, 2016) is designed to help humans understand the decisions of "black-box" models, for example, neural networks, gradient boosting machines, and other ensemble models. These models are accurate, but their complexity is often difficult. LIME helps us predict and identify critical characteristics that influence student dropout. Specifically, LIME starts by selecting a data engine and then creates new sample data by perturbing its specifications. LIME then observes how changing these particular images affects the model's expectations. Based on these surveys, LIME builds a simpler, more understandable model, typically a linear model or decision tree, to simulate the complex model's behavior around the data origin.

3. RESULTS AND DISCUSSION

3.1. Environmental settings and metrics for evaluation

We performed this study using Google Colab and MacBook Air M1 environments. MacBook Air M1 has an Apple Silicon M1 processor with eight cores, 8G RAM, and 1 SSD with 256GB. Google Colab provides 12.67 GB RAM and a virtual machine to support primary and moderate machine learning calculations. We use Google Colab's GPUs to speed up model training and evaluate their performance on test datasets.

For this study, we have chosen to evaluate the model performance in metrics such as Accuracy, Precision,

Recall and F1-Score. The goal here is to verify the validity and reliability of these predictions generated by the model, increasing the likelihood with which we can predict earlier enough in order to intervene. The specific metrics of interest to us are True Positive (TP), the number of correctly identified dropout students, True Negative (TN), the number of correctly identified non-dropout students, False Positive (FP), the number of non-dropout students incorrectly classified as dropouts, and False Negative (FN), or the number of dropout students wrongly designated as non-dropouts. Accuracy (Equation 1) works by dividing the number of correct predictions by the total number of predictions made by the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision (Equation 2): It is the number of Positives the model predicted as Positives (TP) out of all instances model predicted to be positive (TP + FP).

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall (Equation 3) is the number of points that are True Positive (TP)/ Total No. of Positive points (TP + FN). Recall is a fundamental metric as it helps us to see how the model performs in terms of not missing True Positives.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1 - score (Equation 4) unifies the Precision and Recall of the classifier together into a single number by taking their harmonic mean (Hamonic Mean).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

3.2. Dropout prediction results with classical machine learning

The prediction results of the "Dropout" class of each model are shown in Table 2 as Mean ± Standard deviation. The standard deviation represents the variation of the mean. RF shows the lowest performance in accuracy (77.3%) and Precision (81%) and has the lowest recall (76%) among these models, and has a lower F1-score (78.3%) compared to the other three models (Figure 2). RF predicted dropout class with the lowest precision and poorly on the other classes due to its low recall. XGBoost showed the highest performance in accuracy (78.5%) and precision (82.3%), but recall (76.3%) is close to the RF model and relatively low. XGBoost and LightGBM exhibited similar metrics of Accuracy (78.2%), Precision (81.3%), Recall (76.6%), and F1 Score (78.8%). LightGBM showed the most minor standard deviation of recall (3.8%), and CatBoost has the most minor standard deviation of accuracy (1.5%). Both models have a minor F1 Score standard deviation (2.8%) (Table 2). LightGBM and CatBoost were the best performers with the highest recall and F1 Score and suit imbalanced data.

Table 2. Prediction Results for "Dropout" Class of Various Algorithms with Standard Deviation

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.773 ± 0.018	0.810 ± 0.053	0.760 ± 0.046	0.783 ± 0.034
XGBoost	0.785 ± 0.023	0.823 ± 0.042	0.763 ± 0.051	0.790 ± 0.037
LightGBM	0.782 ± 0.021	0.813 ± 0.044	0.766 ± 0.038	0.788 ± 0.028
CatBoost	0.776 ± 0.015	0.812 ± 0.042	0.769 ± 0.045	0.789 ± 0.028

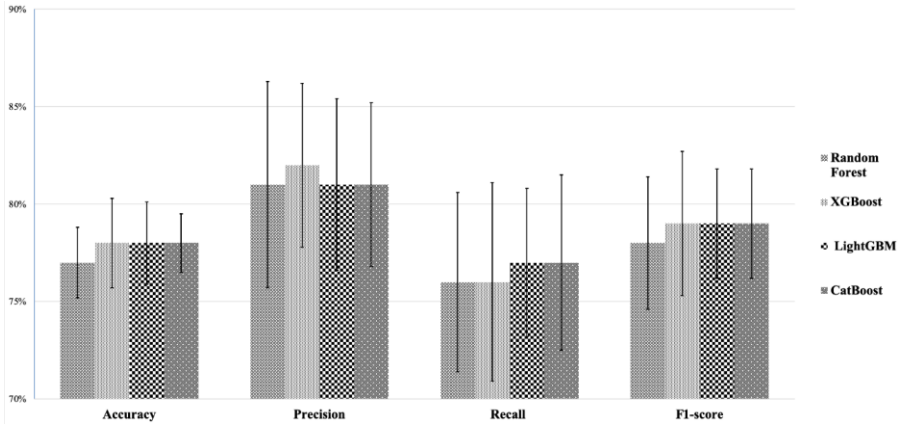

Figure 2. A comparison of "Dropout" prediction of various algorithms with Standard Deviation

Our results show that LightGBM and CatBoost would be the best choices due to their highest F1-score and recall, followed by the XGBoost model. The model quality has been thoroughly tested. The fact that the models have similar performance shows that they can effectively address this issue. Slight differences may reflect that the models perform well under current conditions. However, the findings should be confirmed with different random seeds to obtain stability and robustness or with more training datasets.

3.3. Analysis with LIME on important factors affecting dropout situations

An explanation example is described to understand the LIME result: The left graph in Figure 3(a) shows the confidence interval describing 75% dropout whereas only 25% of non-dropout classes (Graduate the center graph shows the feature importance score with "curriculum units 2nd sem. (approved)" having a 6% value, followed by "curricular units 1st sem. (approved)" and "curricular units 1st sem. (grade)" with 4% and "age at enrollment" and "scholarship holder" both with 2%. The right graph shows the top five features and their respective values. The features highlighted in orange contribute to the dropout class, and those in blue contribute to the non-dropout class.

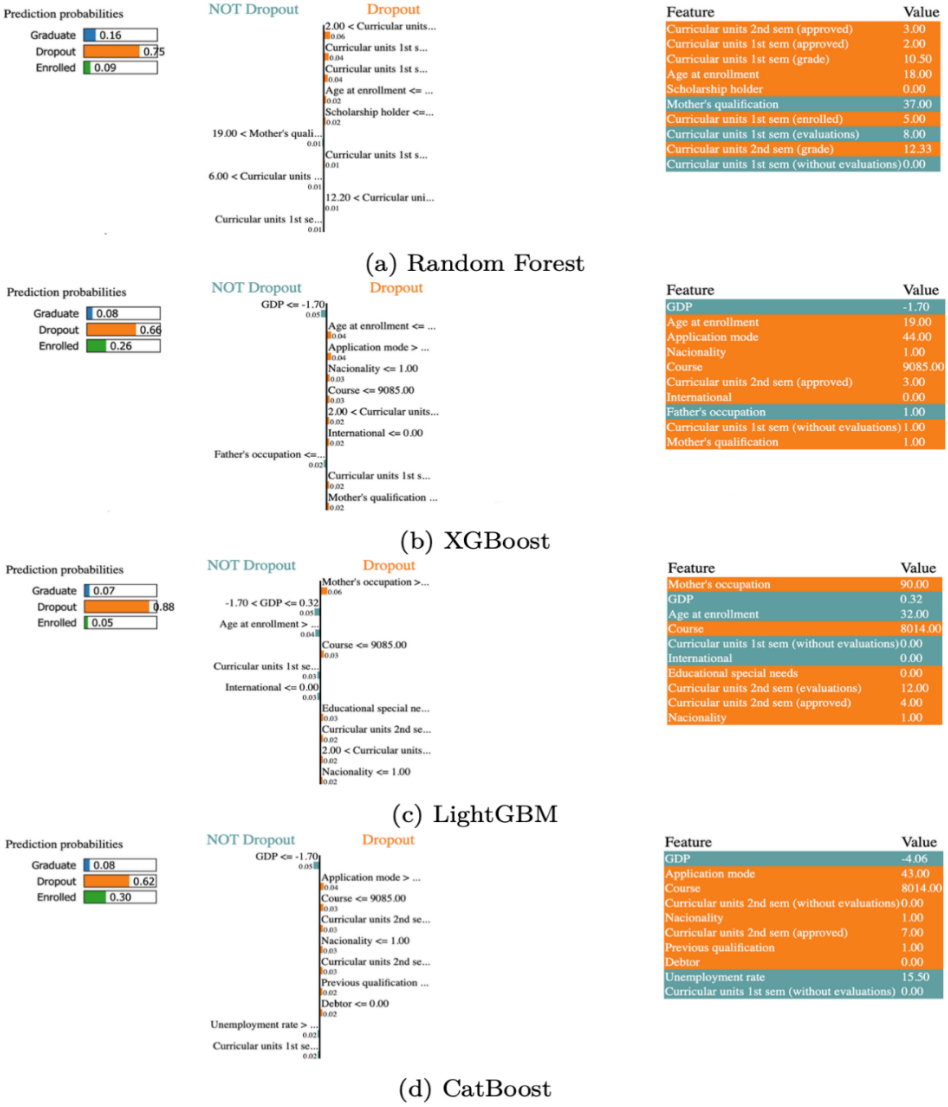

Figure 3. An illustration of explanations on the predictions of various models

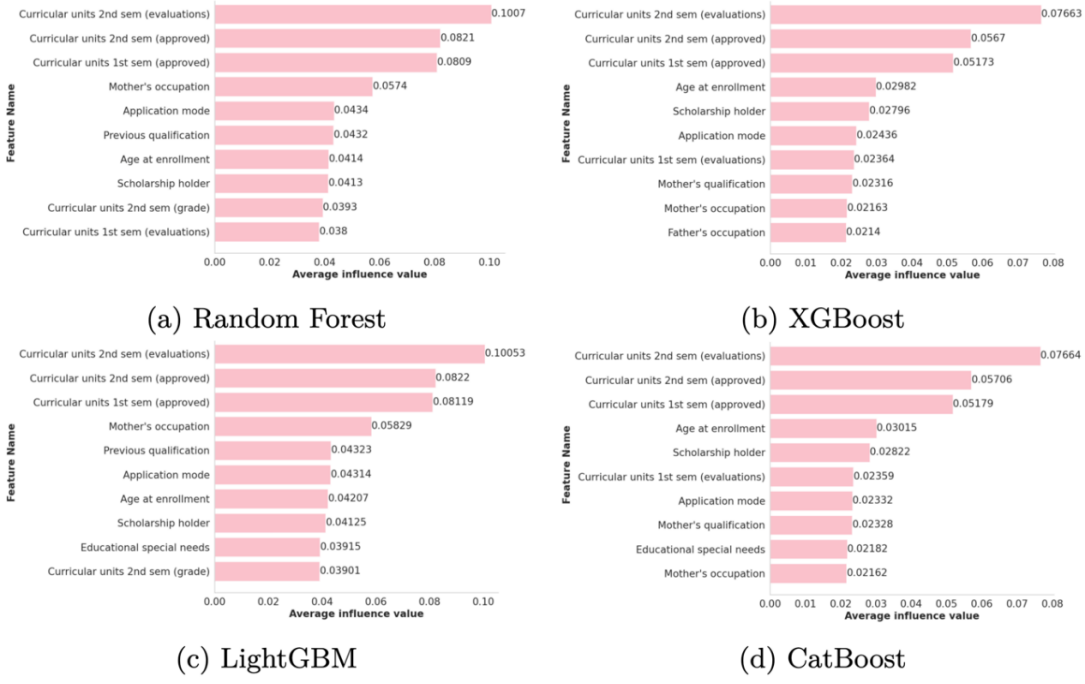

Figure 4. Top 10 Important Features that Affect the Dropout Decision of machine learning models

XGBoost and CatBoost display lower probabilities of dropout prediction (66% and 62% respectively) compared to RF with 75% and LightGBM with 88%. “Curricular units 2nd sem. (approved)” was found in the four models, confirming its importance. The features found in RF are mainly like the original findings in (Realinho et al., 2022) with five features including "curricular units 1st and 2 sem. (credited), (enrolled), (evaluations) and evaluated". “Nationality, course” features appear in XGBoost, LightGBM, and CatBoost. “Age at enrollement” appears in RF and XGBoost models with 4% and 2% values, respectively. The “Mother’s occupation” feature appears in LightGBM with the highest value of 6%.

Most students in the dataset come from Portugal, and nationality correlates with international status. The students without a scholarship occupy 39% in dropout class (Realinho et al., 2022). This indicates that financial aid has a critical impact on dropout classes. The mean age was 23 and strongly correlates with the curriculum units in general (Figure 1). Courses also have a strong correlation with the curriculum units.

3.4. Comparison and Discussion

Table 3 compares the results obtained from this study and those in the original one (Realinho et al., 2022). The difference between our research and previous research with the same dataset (Realinho et al., 2022) is the combination of LIME instead of PFI with machine learning models RF, LightGBM, XGBoost, and CatBoost. PFI provides an overview of the essential features of all machine learning models in (Realinho et al., 2022), and LIME provides a detailed and local interpretation of each specific prediction.

LIME enables the detection of essential features in exceptional cases, whereas PFI may miss them by focusing on global assessments. LIME has identified essential features such as "curricular units 2nd sem. (evaluations)", "mother’s occupation", "application mode", "age at enrollment" and "scholarship holder" for each specific prediction. In contrast, PFI identifies "curricular units second sem. (approved)", "curricular units 1st sem. (approved)", "curricular units 2nd sem. (grade)", "course" and "tuition fees up to date" as crucial features for the entire model.

LIME has many outstanding advantages, such as providing detailed and easy-to-understand explanations at each specific data point and helping end users and experts easily understand and trust the model's predictions. These are significant advances for LIME over PFI, as PFI only focuses on global assessment and may ignore important influencing factors in individual predictions. Therefore, compared with the original research (Realinho et al.,

2022), our results provide a more detailed intuitive explanation, local explanation, black-box models explanation, and application features that allow for personalized interventions and, thus, more practical applications in education.

The highest algorithm in which all measured values are above 76% is the LightGBM and CatBoost models. Seven features were found to significantly impact student dropout rates: curricular units second sem. (approved), curricular units first sem. (approved), curricular units second sem. (evaluations), mother's occupation, application mode, age at enrollment, and scholarship holder.

Table 3. Comparison of Permutation Feature Importance: PFI (Realinho et al., 2022) vs LIME. The features marked with an "x" are those that were identified as important for the entire model.

Feature	PFI (Realinho et al., 2022)	LIME
Curricular units 2nd sem. (approved)	x	x
Curricular units 1st sem. (approved)	x	x
Curricular units 2nd sem. (grade)	x	
Course	x	
Tuition fees up to date	x	
Curricular units 2nd sem. (evaluations)		x
Mother's occupation		x
Application mode		x
Age at enrollment		x
Scholarship holder		x

Financial aid is known as one of the major problems leading to dropout (Dinh-Thanh et al., 2021; Li & Carroll, 2020; Núñez-Hernández & Buele, 2023; Nurmalitasari et al., 2023). Age was identified by (Dinh-Thanh et al., 2021) and (Moreira Da Silva et al., 2022) as an essential feature in evaluating the course's completion. Grades is identified as the primary factor for failure in university, resulting in dropout. In an international environment, non-native-speaking students often face a language barrier, resulting in low performance. A recent study shows that non-native English-speaking students participated less in collaborative learning and co-curricular activities and obtained smaller gains in critical thinking skills (Liu et al., 2021). However, the majority of students in this dataset were Portuguese. About 25% of the students drop out due to the feature "mother's occupation". Mother's occupation may refer to single student mothers, often facing financial and time-related obstacles (Gault & Cruse, n.d.; Kruvelis et al., 2017).

The distribution of these features is skewed. For instance, most students are from Portugal, and the number of scholarship holders is unequal in the population. This impacts the ML models, such as overfitting the majority class, misleading feature importance, and metrics performance. To improve LIME results, more analys is needed, including balancing the data and running LIME multiple times on various datasets sub-datasets to check the consensus of feature importance. Educators and policymakers can then use these critical factors to develop effective strategies to address the dropout problem. The results will help identify the risk factors that influence students' decisions to leave, allowing for effective interventions and support to reduce dropout rates and improve learning outcomes.

4. CONCLUSION

We applied Local LIME to four machine-learning models: RF, CatBoost, XGBoost, and LightGBM. The LightGBM and CatBoost models performed the best, achieving metric values exceeding 76%. This research shows the effectiveness of LIME in providing details; it helps better understand the model's predictions for each individual, which can be crucial in practical applications such as intervention and student support. The results obtained from LIME are significant in predicting student dropout. "Curricular units 2nd sem. (approved)" and "Curricular units 1st sem. (approved)" are two features similar to the original research. Features affecting dropping out that were not found in the original study included "curricular units 2nd sem. (evaluations)", "mother's occupation", "application mode", "age at enrollment", and "scholarship holder".

By identifying and understanding the specific factors that influence dropout prediction, educational institutions can develop targeted interventions to support at-risk students, reducing dropout rates and improving retention, which ultimately leads to better educational outcomes.

This study focuses on a specific educational institution to allow for comparison with previous studies using the same dataset and models, albeit with different implementation methods, while also providing insights into the context of dropout scenarios. However, this study has not yet diversified its dataset or utilized individual machine learning models, which introduces the risk of overfitting. future work should aim to expand the dataset and enhance evaluation metrics by incorporating diverse data sources, combining models, and applying regularization techniques to mitigate overfitting risk. Additionally, extending the study by integrating deep learning models could offer a more comprehensive comparison with traditional methods.

REFERENCES

Belyadi, H., & Haghighat, A. (2021). Supervised learning. In *Machine Learning Guide for Oil and Gas Using Python* (pp. 169–295). Elsevier. https://doi.org/10.1016/B978-0-12-821929-4.00004-4

Dinh-Thanh, N., Thanh-Hai, N., & Thi-Ngoc-Diem, P. (2021). Forecasting and Analyzing the Risk of Dropping Out of High School Students in Ca Mau Province. In T. K. Dang, J. Küng, T. M. Chung, & M. Takizawa (Eds.), *Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications* (Vol. 1500, pp. 224–237). Springer Singapore. https://doi.org/10.1007/978-981-16-8062-5_15

Gault, B., & Cruse, L. R. (n.d.). *Investing in Single Mothers' Higher Education: Higher education*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Kruvelis, M., Cruse, L. R., & Gault, B. (2017). *Single mothers in college: Growing enrollment, financial challenges, and the benefits of attainment*. Briefing Paper #C460. Institute for Women's Policy Research. https://eric.ed.gov/?id=ED612464

Li, I. W., & Carroll, D. R. (2020). Factors influencing dropout and academic performance: An Australian higher education equity perspective. *Journal of Higher Education Policy and Management*, *42*(1), 14–30. https://doi.org/10.1080/1360080X.2019.1649993

Liu, J., Hu, S., & Pascarella, E. T. (2021). Are non-native English speaking students disadvantaged in college experiences and cognitive outcomes? *Journal of Diversity in Higher Education*, *14*(3), 398–407. https://doi.org/10.1037/dhe0000164

Moreira Da Silva, D. E., Solteiro Pires, E. J., Reis, A., De Moura Oliveira, P. B., & Barroso, J. (2022). Forecasting students dropout: A UTAD University Study. *Future Internet*, *14*(3), 76. https://doi.org/10.3390/fi14030076

Núñez-Hernández, C., & Buele, J. (2023). Factors Influencing university dropout in distance learning: A case study. *Journal of Higher Education Theory and Practice*, *23*(14). https://doi.org/10.33423/jhetp.v23i14.6379

Nurmalitasari, Awang Long, Z., & Faizuddin Mohd Noor, M. (2023). Factors influencing dropout students in higher education. *Education Research International*, *2023*, 1–13. https://doi.org/10.1155/2023/7704142

Oreopoulos, P., & Ford, R. (2019). Keeping college options open: A field experiment to help all high school seniors through the college application process. *Journal of Policy Analysis and Management*, *38*(2), 426–454. https://doi.org/10.1002/pam.22115

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2021). Predict students' dropout and academic success (Version 1.0) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.5777340

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, *7*(11), 146. https://doi.org/10.3390/data7110146

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1

Singh, S., & Guestrin, C. (2016). "Why Should I trust you?": Explaining the predictions of any classifier (arXiv:1602.04938). arXiv. http://arxiv.org/abs/1602.04938

Song, Z., Sung, S.-H., Park, D.-M., & Park, B.-K. (2023). All-year dropout prediction modeling and analysis for university students. *Applied Sciences*, *13*(2), 1143. https://doi.org/10.3390/app13021143

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society*, *76*, 102474. https://doi.org/10.1016/j.techsoc.2024.102474