Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

# Methods of Balancing Model Explainability and Performance in Identifying At-Risk Students

**Tiffany T.Y. HSU[a], Brendan Flanagan[b] & Owen H.T. LU[a*]**
[a]*International College of Innovation, National Chengchi University, Taiwan*
[b]*Center for Innovative Research and Education in Data Science, Kyoto University, Japan*
*owen.lu.academic@gmail.com

**Abstract:** This study will explore and experiment with various combinations of methods to handle data imbalance in order to address the common issue of insufficient minority samples in at-risk student prediction. Additionally, we will examine the purpose of applying computer tools to educational issues and emphasize the necessity of adhering to models with high transparency and explainability, ensuring that the decision-making process can be transparent and comprehensive in the context of learning analytics. After comparing model performance, we selected the logistic regression model combined with correlation analysis and threshold adjustment, which showed outstanding performance in UAR, G-means, and other evaluation metrics. We will analyze the reasons behind students' academic performance based on the feature importance ranking from the model, thereby establishing a high-performance and high-transparency benchmark model for the LBLS593 dataset.

**Keywords**: Learning analytics, Imbalanced data, At-risk student prediction, Correlation analysis, Model explainability, Trustworthiness

## 1. Introduction

Identifying at-risk students is a recurring issue in learning analytics(LA). To recognize the behavioral tendencies of at-risk students in the early stage and assist teachers in developing intervention strategies, we continuously collected students' learning data on different learning platforms. This year, a dataset named LBLS593 has been released and is being used for research in learning analytics.

This study is going to establish a model using the data collected in 2023 of 467 students as a training dataset and take the newly updated data from 126 students as a predicting target by comparing the performance of different machine learning models and data preprocessing methods, establishing a new accuracy benchmark, and finally identifying the features with more significant influence on the final prediction. Different from before, we have realized the deficiencies in utilizing explainable AI methods (XAI). Despite the high applicability of XAI models such as SHAP and Lime, XAI methods can provide unfaithful and insufficient explanations, failing to clarify the black box model's computations (Rudin, 2019). For instance, explaining the SHAP model can be highly complicated while tackling a high-dimensional dataset. With an increasing number of features, the complexity of the distribution of Shapley value is rising simultaneously, leading to difficulty for teachers in comprehending and making use of the result (Ortigossa et al., 2024). Likewise, it is notable that the XAI model exhibits a certain amount of instabilities even if the base model is stable (Dai et al., 2022). If the explanation provided by the XAI method fluctuates and differs over multiple runs, the reliability is questionable and potentially causes partial decisions or treatments.

Besides the potential issues that may arise from the computational process of the XAI model, with higher demands of transparency in the model decision process, accuracy and recall rate are not the only indicators to evaluate the models' performance. The explainability of the machine learning models themselves is gaining more attention. Explanations of models are not merely a matter of ethics and fairness but also necessary information for teachers to

formulate intervention strategies, especially in education that emphasizes context and individuality. Therefore, to maintain the explainability of the model, we aim to use only simple machine learning models like logistic regression, decision trees, and other models based on simple algorithms instead of deep learning models like support vector machines and neural networks. In this study, we will answer the following two research questions:

> **RQ1:** To balance the explainability and accuracy of the model, what method can be applied in the context of learning analytics?
> **RQ2:** What are the key factors that influence students' performance?

## 2. Literature Review

Several educational institutions collect students' behavioural data on digital learning platforms, aiming to help the teaching faculties to track students' situations and provide on-time assistance, especially in computer science education, in which quantitative data collection is more accessible thanks to the emergence of various online learning platforms. That is also why learning analytics applications primarily focus on computer programming classes or computer science subjects. Learning datasets used by research like this include features like assignment submission time (Falkner et al., 2012), learning activities during weekends, class attendance, and how long they spend on the learning platform. Some datasets even contain data outside of class, such as travel distance to the university and grades of other courses (Azcona et al., 2019). Collecting these data aims to discover subtle or apparent factors that affect students' academic performance. The features in LBLS datasets include learning activities and numerous self-regulated learning and learning strategy-inventory measurement results reflecting students' attitudes, motivations, and learning styles(Lu et al., 2022), which are proven to impact the learning process significantly by learning theories (Rößling et al., 2008)

When it comes to defining at-risk students, we have drawn on a wide range of related research and have categorized them into two types. The first type is based on behaviors. For instance, according to Falkner et al. (2012), at-risk students are those with poor assignment submission status. In Azcona et al. (2019) and Al-Shabandar et al. (2019), successful behavioural patterns have been established by tracking past behavioural trajectories. Students who deviate from these expected study patterns are identified as at-risk. The second type is based on score criteria, such as low GPA or grades in mid-term and final exams. In our study, we will adopt the latter approach and define at-risk students as those whose final scores fall within the bottom 3% of the class due to the deficiency of legacy data.

While exploring the prediction of at-risk students, we discovered a frequent issue of imbalanced datasets. Since at-risk students are normally anomalous within the course, even though the accuracy of classifiers can be very high, the low recall rate indicates that the models fail to detect the anomalies. Haixiang et al. (2017) have provided an overview of frequently applied methods like re-sampling, feature extraction, cost-sensitive, ensemble methods, and algorithmic classifier modifications, and elucidate the features of dataset and scenarios for which these methods are most appropriately suited. The combined use of these methods can as well be discovered in learning analytics. Barros et al. (2019) utilized balanced bagging and tested various performance evaluation metrics to address the fact that students-dropout are significantly fewer than those who persist in their studies in a research of predicting student dropout. To deal with a similar issue Thammasiri et al. (2014) compare the efficiency of multiple re-sampling methods; Hlosta et al. (2017) however, used a class weight adjustment method the feature. Owing to the dynamic feature of their dataset, students' behavior and data distribution are constantly changing, indicating an one-off re-sampling methods for static dataset is not applicable in this scenario. in Severson et al. (2007) with their dataset contained the students' behavioural data on virtual leanring platform that has a similar dynamic features, the research applied threshold adjustment methods to influence the output, intervening during the model predicting stage. The aforementioned studies have affirmed the the necessity of selecting appropriate strategies based on the characteristics of imbalanced datasets. We classify these methods into three categories based on the different stages of the experiment

where the techniques are applied: data preprocessing, model training, imbalance handling, and model post-processing. The methods used in different stage should be confirmed after identifying the characteristics of our dataset.

The purpose of at-risk students' prediction is not simply labelling. Instead, it is to let teachers provide on-time assistance when students encounter issues in their learning progress. In advance of any intervention, the required information is the explanation of the prediction. 'Any generalization or theory constructed without deep understanding, not grounded in the concrete and particular, is vacuous' (Birhane, 2021). The over-generalization of the machine learning process can potentially result in issues of students' individuality being overlooked, along with bias and discrimination (Scholes, 2016). We assert that the explainability and transparency in at-risk student prediction must not be compromised. Therefore, choosing the model and imbalanced handling methods will be the primary consideration in determining whether the method aligns with the context and needs of at-risk student prediction. For example, we will avoid using feature extraction techniques that can reduce explainability, such as PCA (Principal components analysis). Substantially, teachers need to understand the key factors influencing student performance; therefore, we prioritize interpretable models over complex deep learning models. Also, we avoid oversampling methods like SMOTE (Synthesized Minority Oversampling Technique), which addresses the imbalance by generating synthetic data points. This approach ensures that the results reflect students' learning conditions based on real-world data in the model interpretation stage.

## *3. Methods*

### *3.1 LBLS593*

This study utilizes a dataset named LBLS593. The dataset was collected from a long-running university programming course and includes a total of 593 students across 12 classes. Each dataset contains 18 weeks of students' BookRoll e-book reading behaviors, VisCode programming editing behaviors, self-regulated learning strategies, language learning strategies, and their final exam scores.

### *3.2 Data Balancing*

After missing value cleaning and normalization, 242 students' data are in the training dataset. Only nine students are labelled as 'at-risk', with scores ranging from 3 to 76 The table below presents the model performance without any imbalance handling methods involved. Despite both models achieving high accuracy rates of 0.938 and 0.800, the F1-scores for class 1 (at-risk) are 0.33 and 0.00, respectively. Logistic Regression demonstrated that only 50% of the predictions for the minority class were correct, and it captured merely 25% of the actual minority class samples. Contrarily, Decision Tree struggled significantly with the minority class, showing no successful identification of 'at-risk' students (precision and recall both 0.00), indicating that none of the actual minority class samples were correctly classified.

Table 1. *Performance Comparison of Logistic Regression and Decision Tree Models Without Data Balancing*

| Evaluation Metrics | Logistic Regression | Decision Tree |
|---|---|---|
| Accuracy | 0.938 | 0.800 |
| F1 Score | 0.33 | 0.000 |
| Precision (Class 0) | 0.95 | 0.93 |
| Recall (Class 0) | 0.98 | 0.85 |
| F1 Score (Class 0) | 0.97 | 0.89 |
| Precision (Class 1) | 0.50 | 0.00 |
| Recall (Class 1) | 0.25 | 0.00 |
| F1 Score (Class 1) | 0.33 | 0.00 |

To improve this situation, we will employ various method combinations, and determining the most effective combination requires empirical validation through experimentation. We classify methods for handling data imbalance into three categories based on the different stages of the experiment where the techniques are applied: data preprocessing, model training, imbalance handling, and model post-processing. We have noticed that some method combinations may have the issue of overlapped work; for example, the combined use of cost-sensitive and class adjustment results in certain feature weights being repeatedly reinforced or weakened. Therefore, eleven method combinations were established after assessing each method's principles and applicable stages.

Table 2. *Categorization of Imbalance Handling Techniques Based on Application Stages*

| Data Preprocessing | Model Training | Imbalance Handling | Model Post-Processing |
|---|---|---|---|
| • Down-Sampling<br>• Correlation Analysis | • Lasso-Regression<br>• Random Forest Classifier<br>• Logistic Regression<br>• Decision Tree Classifier | Class Adjustment | Threshold Adjustment |

Table 3. *Classification of Model Combinations by Imbalance Handling Strategy*

| | Categorization | Description |
|---|---|---|
| 1 | Based on Sampling Strategy | Down-sampling + Random Forest Classifier |
| 2 | | Down-sampling + Lasso Regression |
| 3 | | Down-sampling + Logistic Regression |
| 4 | | Down-sampling + Decision Tree Classifier |
| 5 | Based on Feature Selection or Correlation Analysis (Pearson correlation coefficient) | Correlation Analysis + Random Forest Classifier |
| 6 | | Correlation Analysis + Logistic Regression |
| 7 | | Correlation Analysis + Lasso Regression |
| 8 | | Correlation Analysis + Logistic Regression + Threshold Adjustment |
| 9 | Based on Class Adjustment & Threshold Adjustment | Class Adjustment + Random Forest Classifier + Threshold Adjustment |
| 10 | | Class Adjustment + Logistic Regression + Threshold Adjustment |
| 11 | | Class Adjustment + Decision Tree Classifier + Threshold Adjustment |

Regarding the model performance evaluation, in addition to accuracy, recall rate and F1 score, we will use evaluation metrics which are more sensitive to imbalance datasets, such as Geometric Mean (G-mean) and Unweighted Average Recall (UAR), to avoid accuracy paradox and capture the model's sensitivity to the minority class (Barros et al., 2019; Chachoui et al., 2024). UAR is primarily used in multiclass classification problems. It is calculated by averaging the recall rates across all classes, without considering the number of samples in each class, making UAR useful for handling class imbalance issues in binary classification problems as well. On the other hand, G-means combines sensitivity (recall) and specificity (true negative rate). Both offer robust metrics for evaluating model performance, especially in scenarios involving class imbalance.

### *3.3 Model Explanation*

The model explanation will be demonstrated using the rank of feature importance. By extracting the coefficient of every feature, we can see the strength and direction of feature influence. This is also the reason for using simple machine learning models; during the interpretation phase of predictions, they can most accurately and straightforwardly reflect the reasons behind the model's decisions. This approach provides insights into the model's overall feature influence and allows for understanding individual data points, enabling teachers to grasp and comprehend students' learning conditions more easily.

## 4. Results

### *4.1 Reply RQ1*

The table below presents the performance evaluation of various method combinations in handling imbalanced data, focusing on metrics such as Accuracy, F1 Score, UAR, and G-means. The "Correlation Analysis + Logistic Regression + Threshold Adjustment" combination outperformed others on all the metrics, achieving the highest F1 Score (0.75), UAR (0.87), and G-means (0.86), showing the best performance in balancing sensitivity and specificity. Conversely, "Down-sampling + Lasso Regression" and "Correlation Analysis + Lasso Regression" performed poorly, with F1 Scores and G-means of 0.0, indicating a failure to recognize most minority classes. The analysis suggests that feature selection combined with threshold adjustment, particularly when paired with logistic regression, performs more consistently and effectively across all metrics. We will use this combination as the model for subsequent interpretation.

Table 4. *Performance Metrics Comparison Across Different Model Combinations for Imbalance Handling*

| Combination | Accuracy | F1 Score | UAR | G-means |
|---|---|---|---|---|
| Down-sampling + Random Forest Classifier | 0.95 | 0.57 | 0.74 | 0.70 |
| Down-sampling + Lasso Regression | 0.94 | 0.00 | 0.50 | 0.00 |
| Down-sampling + Logistic Regression | 0.94 | 0.50 | 0.73 | 0.70 |
| Down-sampling + Decision Tree Classifier | 0.64 | 0.08 | 0.46 | 0.41 |
| Correlation Analysis + Random Forest Classifier | 0.91 | 0.25 | 0.60 | 0.49 |
| Correlation Analysis + Logistic Regression | 0.97 | 0.67 | 0.75 | 0.70 |
| Correlation Analysis + Lasso Regression | 0.94 | 0.00 | 0.50 | 0.00 |
| Correlation Analysis + Logistic Regression + Threshold Adjustment | 0.97 | 0.75 | 0.87 | 0.86 |
| Class Adjustment + Random Forest Classifier + Threshold Adjustment | 0.89 | 0.22 | 0.59 | 0.48 |
| Class Adjustment + Logistic Regression + Threshold Adjustment | 0.92 | 0.44 | 0.73 | 0.69 |
| Class Adjustment + Decision Tree Classifier + Threshold Adjustment | 0.92 | 0.29 | 0.61 | 0.49 |

In addition to the four metrics, we use Receiver Operating Characteristic Curve (ROC) curves to help us get a more intuitive and easier way to interpret the model performance. ROC curve is a graphical tool for evaluating the performance of binary classification models, illustrating the model's trade-off between True Positive Rate and False Positive Rate under different thresholds. The diagonal line in the graph represents the performance of a random

guessing model, and the Area Under Curve (AUC) represents the model's performance. A figure closer to 1 indicates that the model can flawlessly execute classification work. On the other hand, closer to 0 means the model's classification ability is worse than random guessing. The ‘Correlation Analysis + Logistic Regression + Threshold Adjustment’ combination reached the highest AUC of 0.99, and its curve is very close to the top-left corner at 1.0, showing that it can perform accurate classification tasks, which aligns with the metrics on the above table.

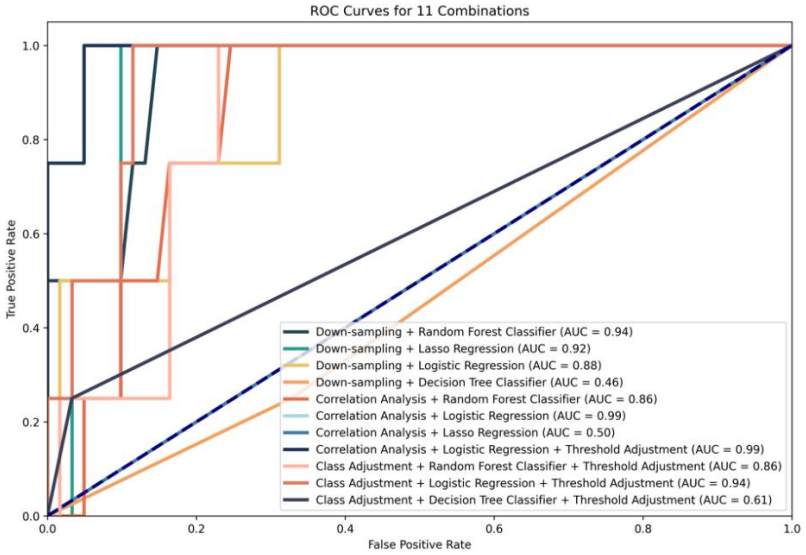


*Figure 1.* ROC curves for eleven combinations

## *4.2 Reply RQ2*

After using the 'Correlation Analysis + Logistic Regression + Threshold Adjustment' combination, we listed the top ten features that have the greatest impact on the prediction results in the logistic regression model. These include srl_m_29 (SRL Measurement: 'I take tests thinking of the consequences of failing.'), PREV (Students' BookRoll activities: 'Went to the previous page.'), srl_m_30 (SRL Measurement: 'I have an uneasy, upset feeling when I take an exam.'), indicating that students' anxiety and frustration regarding exams may be significant factors influencing their academic performance. Additionally, the frequency of repeatedly returning to the previous page in BookRoll activities could indicate the frequency of review, which may also have a substantial impact on their learning outcomes.

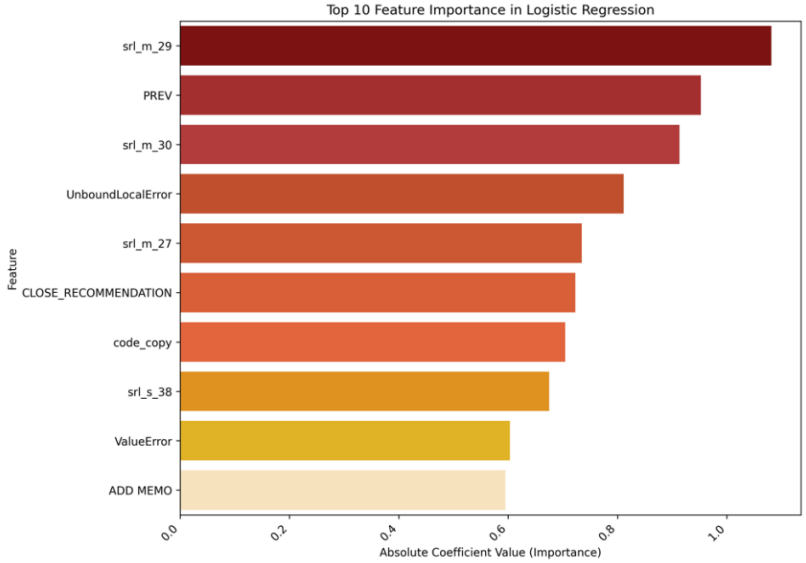


*Figure 2.* Top 10 feature importance in logistic regression model

## 5. Conclusion

The dilemma between model accuracy and explainability is manifested in the issue of at-risk student prediction. Due to the lack of anomaly samples, the model's explainability is often compromised in pursuit of accuracy. However, this approach does not align with the context of educational issues, especially when using computer tools for human-centred issues; the entire process should align with human understanding since the establishment of a prediction model does not ask for high accuracy. Instead, it allows teachers to understand the student's situation. A model with interpretability is essential in educational practice and necessary to ensure that students' individuality is noticed due to the overgeneralization often seen in machine learning. Moreover, the widely held belief that "accurate models must be highly complex" has been challenged by numerous studies. We have demonstrated in this study that even with simple models, significant performance improvements can be achieved through appropriately handling imbalanced data, establishing a new benchmark for the LBLS593 dataset: a highly effective, highly interpretable, and a model low in complexity.

## References


Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2019). Detecting at-risk students with early interventions using machine learning techniques. *IEEE Access*, *7*, 149464-149478.

Azcona, D., Hsiao, I.-H., & Smeaton, A. F. (2019). Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. *User Modeling and User-Adapted Interaction*, *29*, 759-788.

Barros, T. M., Souza Neto, P. A., Silva, I., & Guedes, L. A. (2019). Predictive models for imbalanced data: A school dropout perspective. *Education Sciences*, *9*(4), 275.

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns*, *2*(2).

Chachoui, Y., Azizi, N., Hotte, R., & Bensebaa, T. (2024). Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning. *Computers and Education: Artificial Intelligence*, *6*, 100222.

Falkner, N. J., & Falkner, K. E. (2012). A fast measure for identifying at-risk students in computer science. Proceedings of the ninth annual international conference on International computing education research,

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, *73*, 220-239.

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: early identification of at-risk students without models based on legacy data. Proceedings of the seventh international learning analytics & knowledge conference,

Lu, O., Huang, A., Flanagan, B., Ogata, H., & Yang, S. (2022). A quality data set for data challenge: featuring 160 students' learning behaviors and learning strategies in a programming course. the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education,

Rößling, G., Joy, M., Moreno, A., Radenski, A., Malmi, L., Kerren, A., Naps, T., Ross, R. J., Clancy, M., & Korhonen, A. (2008). Enhancing learning management systems to better support computer science education. *ACM SIGCSE Bulletin*, *40*(4), 142-166.

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

Scholes, V. (2016). The ethics of using learning analytics to categorize students on risk. *Educational Technology Research and Development*, *64*(5), 939-955.

Severson, H. H., Walker, H. M., Hope-Doolittle, J., Kratochwill, T. R., & Gresham, F. M. (2007). Proactive, early screening to detect behaviorally at-risk students: Issues, approaches, emerging innovations, and professional practices. *Journal of School Psychology*, *45*(2), 193-223.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert systems with applications*, *41*(2), 321-330.