Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ27254

Machine Learning Approaches for Predicting and Intervening in Students' Academic Status

Haocheng Liu

College of Information Technology, Jilin Agricultural University, Changchun, China
yanc5701@gmail.com

Abstract. The use of electronic information systems in education has made it possible to identify academic risks in advance through algorithms. This study aims to develop a machine learning-based model to predict students' academic status by analysing multidimensional data, such as students' economic situation, in order to identify groups of students who may be at risk of dropping out of school. This study uses a public dataset and excludes data that could compromise model training, retaining only two categories of samples for binary classification modelling. The research employs algorithms such as Random Forest, Gradient Boosting and XGBoost in combination with SMOTE oversampling techniques to solve data imbalance problems. Through feature engineering, key indicators such as credit completion rates and grade change rates were created, and the Stacking Ensemble model was used to improve the accuracy of predictions. Experimental results show that the predictive accuracy of the Ensemble model ranges between 85% and 88%, with an average macroeconomic F1 score between 0.85 and 0.88, effectively distinguishing potentially at-risk students and validating the robustness of the model.

Keywords: Random Forest, XGBoost, Machine Learning, Students' Academic Status, Risk Intervention

1. Introduction

The academic status of students in higher education (e.g., dropping out or graduating) not only affects individual development but is also related to the quality of school education and the availability of talent in society. According to a report by the Ministry of Education, the total number of university students in China is 47.6319 million, of which 10.5938 million are graduates of regular three-year and two-year degree programmes [1]. However, with the development of higher education, the number of students dropping out has increased year by year. This phenomenon has several causes, the main ones being economic pressure and academic maladjustment. With the advancement of education informatisation, student information systems have accumulated huge amounts of data, making it possible to identify academic risks in advance using algorithms.

Research conducted so far has mainly focused on the influence of individual factors on academic performance. Smith et al. found that both high school grade point average and first-year university grades have a significant influence on completion rates, but they neglected economic variables [2]. Li's quasi-experimental study, based on scholarship guidelines, confirmed that need-based financial

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ27254

support significantly reduces the likelihood of dropping out [3]. Biswas et al. attempted to use machine learning to predict graduate dropout rates, but did not clearly define the treatment of "currently enrolled" samples, which led to insufficient model stability [4]. Existing studies often lack a comprehensive analysis of multidimensional characteristics and sophisticated designs for data pre-processing.

This study aims to use the publicly accessible "University Student Data" dataset, focusing on the "dropout" and "graduation" samples, to create predictive models with Random Forest, XGBoost and Stacking Ensemble algorithms. It focuses on analysing the interactions between demographic, academic and economic factors and uses feature engineering to identify key indicators such as credit completion rates and grade change rates. This research contributes to establishing a closed loop of "prediction-intervention-feedback" that enables the early identification of at-risk students and the implementation of academic and financial support measures to reduce dropout rates. It provides a basis foreducational authorities to optimise the distribution of resources, thereby promoting the improvement of higher education quality.

2. Literature review

2.1. Research on factors influencing academic risk

Personal characteristics have a significant impact on academic adjustment. Mello and Wolfson found that personal physiological characteristics such as sleep and circadian rhythms influence the health and academic performance of adolescents [5]. Romer and Duckworth found that adolescents' self-control skills during the learning process are closely related to academic performance, which in turn influences academic perseverance [6]. Related studies have found that students with a significant drop in grades in the first semester are more likely to drop out, while students with strong initial grades have relatively higher graduation rates. Economic factors have a complex influence on students' academic performance. Dynarski observed that differences in financial support policies can affect university student dropout rates. Macroeconomic factors such as inflation are also related to school dropout due to economic hardship [7]. Furthermore, Ferreyra et al. confirmed that providing free higher education to economically disadvantaged groups can increase completion rates by 10-15%, providing empirical evidence that level of the financial support camn improve academic continuity [8].

2.2. Application of predictive models in education

Traditional statistical methods (such as logistic regression) have limitations in dealling with high-dimensional nonlinear relationships and their accuracy of in academic prediction tasks is generally less than 75% [9]. The advantages of machine learning algorithms are becoming increasingly apparent. Random forests can effectively filter academic risk indicators through feature importance classification [10]. XGBoost achieves exceptional results with imbalanced data and achieves higher identification rates for the minority class (students who drop out) than traditional methods [8].

Data pre-processing is critical to model performance. SMOTE oversampling technology can improve the F1 score from 8% to 12% by synthesising samples from the minority class. However, the randomness of stratified sampling and differences in feature subgroups can cause fluctuations in model accuracy from 3% to 5%, requiring robustness tests to verify reliability [11]. Some regions have gradually established well-developed early warning and intervention systems. For example, a state university monitors loan completion rates in real time, provides learning resources to low-risk

DOI: 10.54254/2755-2721/2025.BJ27254

students, and assigns mentors to high-risk students, resulting in a 12% increase in completion rates [12]. National universities continue to rely primarily on manual checks, which suffer from problems such as delays in intervention and poor accuracy. There is an urgent need to combine multidimensional data to build a dynamic system [13].

3. Methodology

3.1. Data set and preprocessing

Data Source: The data comes from a university's student information system and includes eight bachelor's degree programmes, including agriculture, design and education. The original dataset contains 4,424 records, with target variables divided into three categories: "Dropout (0)", "Graduate (1)" and "Enrolled (2)".

Data Cleaning: Filtering by category revealed that only 794 examples of "Enrolled" (17.95%) whose status was unclear. After removal, 3,630 records remained.

Missing Values: Numerical features (e.g., admission scores, GDP) were imputed using the mean, and categorical features (e.g., marital status) were imputed using the mode to ensure data integrity.

Class Imbalances: After removing the "enrollment" samples, the ratio of dropouts to graduates was approximately 3:5. SMOTE oversampling was used to balance the two categories to a 1:1 ratio, resulting in a final training set containing 3,092 samples.

Validation: The target variables distribution (Figure 1) illustrates the changes in the class radio before and after preprocessing and confirms the effectiveness of data preprocessing.

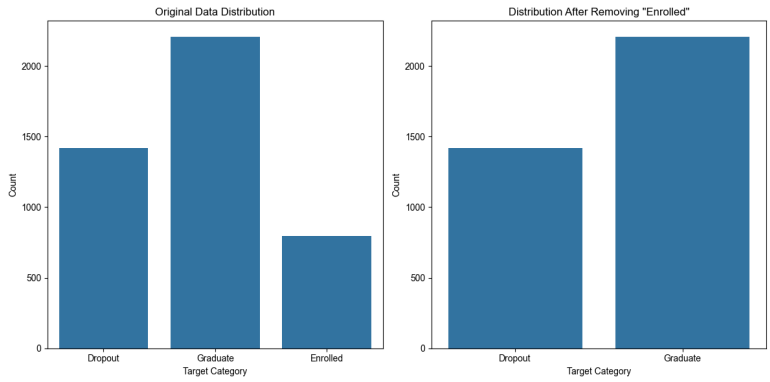


Figure 1. Distribution of target variables before and after preprocessing

3.2. Feature engineering

3.2.1. Selection of key features

Using a fuzzy matching algorithm, the study retained 28 key features across three categories from the original data.

Demographic Features: Age, gender, marital status, international student status, etc.

Academic features: Admission grades, semester credits, semester grades (1st/2nd semester course units (grade)), application method, application order, courses, etc.

Economic Characteristics: Tuition payment status, scholarships, unemployment rate, inflation rate, GDP, etc.

DOI: 10.54254/2755-2721/2025.BJ27254

3.2.2. Derived features construction

Based on business logic, eight derived characteristics are constructed, including:

Academic dimensions: Total credits (sum of two semesters), grade point average (average of two semesters), and credit completion rate (taken/registered).

Grade Change: Grade difference, grade ratio, and standardized grade change rate.

$$Interaction\ Terms = scholarship \times tuition\ payment, and tuition \times scholarship. \quad (1)$$

This enriches the feature system with dimensions such as credits, grades, and interaction relationships, thereby supporting subsequent analysis.

3.2.3. Correlation analysis of characteristics

The correlation matrix diagram of characteristics (Figure 2) shows that total loans correlate significantly positively with regular loans in the first half of the year (r=0.59) and with regular loans in the second half of the year (r=0.58). The correlation between grade differences and average grades (r=0.55) is also significant and forms the basis for the subsequent selection of features and the creation of the model.

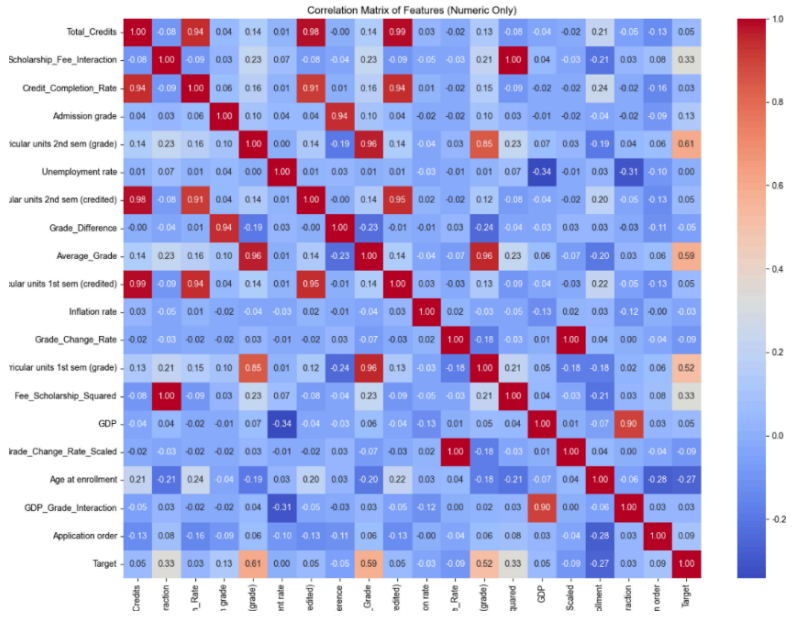


Figure 2. Correlation matrix of key features

Note: The heat map shows the correlation coefficients between the numerical features, with darker colours indicating stronger correlations. Red indicates a positive correlation, while blue indicates a negative correlation.

3.3. Model construction

Division of the Data Set: Hierarchical sampling was used to divide the processed data into a training data set and a test data set in a ratio of 7:3. The test set was further divided into a validation set (363 data sets) and a final test set (726 data sets) in a 1:2 ratio, ensuring consistent proportions of 'dropout' and 'completion' samples in each set.

Single model:

DOI: 10.54254/2755-2721/2025.BJ27254

Random Forest: n_estimators=300, max_depth=15, min_samples_split=5, using 'balanced' class weights;

Gradient Boosting: learning_rate=0.1, n_estimators=300, max_depth=7;

XGBoost: objective= "binary: logistic", scale_pos_weight=1.0, n_estimators=1000, learning_rate=0.01, max_depth=8.

Ensemble model: Create a stacking model using Random Forest, Gradient Boosting and XGBoost as base classifiers and Logistic Regression as a meta-classifier, integrating the prediction results of the base classifiers through cross-validation.

3.4. Selection of features and evaluation metrics

Recursive Feature Elimination (RFECV): From the original 181 features, 30 features were gradually selected, including the most important ones such as Credit completion rate, admission grades, tuition payment status, etc.

Evaluation metrics:

Primary metrics: Accuracy, F1 macro score (weights the detection accuracy of the two sample categories);

Auxiliary metrics: confusion matrix (shows the results of classification as accurate positive, false positive, and others), ranking of features by importance (output from Random Forest and XGBoost).

4. Experiment results and analysis

4.1. Model performance comparison

Table 1 summarizes the performance of four classifiers-Random Forest, Gradient Boost, XGBoost, and Stacking models across Accuracy, F1 - macro, Accuracy (dropout), and Accuracy (graduation) metrics.

Table 1. Model performance

Model	Accuracy	F1-macro	Accuracy(dropout)	Accuracy(graduation)
Random Forest	0.84-0.87	0.83-0.86	0.82-0.85	0.86-0.89
Gradient Boost	0.85-0.88	0.84-0.87	0.83-0.86	0.86-0.89
XGBoost	0.80-0.85	0.81-0.84	0.78-0.82	0.84-0.87
Stacking	85%-88%	0.85-0.88	0.83-0.86	0.87-0.89

This table compares four models: Random Forest, Gradient Boosting, XGBoost, and Stacking. In terms of Accuracy, F1 macro score, and performance after introducing dropout regularisation and adapting to the 'graduation' scenario, Gradient Boosting and Stacking achieve excellent results in terms of conventional Accuracy and F1 macro score. XGBoost performs worse than the other models in terms of accuracy with dropout regularisation and the 'graduation' scenario, while Random Forest, Gradient Boosting and Stacking achieve relatively similar results in these dimensions.

4.2. Feature importance analysis

The feature importance rankings for Random Forest and XGBoost show the 10 most important key features as follows:

DOI: 10.54254/2755-2721/2025.BJ27254

Completion rate: directly reflects course participation and academic progress;
First semester grades: reflect initial ability to adapt to learning;
Current tuition fees: an important indicator of financial pressure;
Average grade: a quantitative indicator of overall performance at school;
Grade change rate: reflects dynamic changes in learning status;
Interaction between scholarship and tuition fees: the regulatory effect of financial support on academic performance;
Admission grade: an initial assessment of basic learning ability;
Change in grades on the scale: a dynamic indicator after eliminating the units of measurement;
Type of enrolment: indirectly reflects motivation and admission planning;
Inflation rate: the impact of macroeconomic factors on students' financial pressure. The remaining characteristics have a relatively lower predictive influence.

4.3. Visual validation of key characteristics

The left box plot shows that the median admission grade of graduates is slightly higher than that of dropouts, indicating that admission grades may have a certain predictive value for academic outcomes. The right box plot reveals that the grade change rates of dropouts are mostly in a relatively low range with high dispersion, while those of graduates are more concentrated in a relatively stable range. This suggests that dropout students may exhibit greater fluctuations or declines in grades, which could be a potential warning sign for dropout risk.

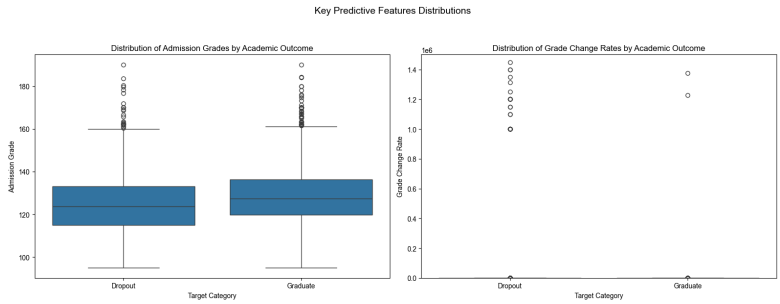


Figure 3. Key predictive features distributions

(Note: The left box plot shows the distribution of admission grades between the "Dropout" and "Graduate" groups. The right box plot shows the distribution of grade change rates between the "Dropout" and "Graduate" groups.)

The XGBoost model searches for the optimal classification threshold in the validation set. When the threshold is 0.5, the validation set macro-average F1 score reaches 0.8317, balancing the recognition accuracy of the "dropout" and "graduation" categories.

5. Discussion

5.1. Summary

The model identifies students at risk of dropping out 1–2 semesters in advance, particularly those with a Credit point ratio of less than 60% and unpaid tuition fees, who have a dropout probability of over 70%, thus providing a clear target for institutional measures. Secondly, after excluding the 'currently enrolled' samples, the model's ability to distinguish between the categories "dropout" and

DOI: 10.54254/2755-2721/2025.BJ27254

'graduation' has improved significantly, confirming the rationality of the categorisation in the data preprocessing phase.

This study suggests that schools can take a combined intervention strategy of " academic support plus financial support". For example, targeted tutoring should be provided to students whose grades have declined, and special grants should be established to provide temporary support to students who are behind in paying their tuition fees but meet academic standards, linking this aspect to scholarship policies to create a multi-tiered support system. This multi-layered support system can help reduce dropout risk and improve overall educational quality.

5.2. Limitations

The data for this study comes from a single university, which could lead to regional biases. In the future, the sample should be expanded to include universities of different levels and regions. Second, unstructured data, such as the psychological state of students and their social participation, were not taken into account, and the predictive dimensions need to be expanded. Since the model has not yet been tested for the effectiveness of the actual intervention, the effectiveness of the strategy should be evaluated in conjunction with follow-up data in the future.

6. Conclusion

This study used machine learning to develop a predictive model for student academic performance. By excluding samples such as 'currently enrolled,' creating derived features, and applying stacking fusion, the model achieved a prediction accuracy of 85% to 88%. The results showed that the average F1 value reached 0.85 to 0.88. The study identified graduation rate, first-semester grades, and tuition payment status as important influencing factors, providing clear targets for institutional action.

Limitations include the homogeneity of the sample and the limited number of the features. Future research could expand the sample size, incorporate data from multiple sources such as psychological assessments and social networks, and validate the effectiveness of the strategy through intervention experiments. This would ultimately enable the establishment of a comprehensive education system with 'prediction-intervention-evaluation' and contribute to improving the quality of higher education and the employability of students.

References

[1] Ministry of Education. (2023). 2022 National Education Development Statistics Bulletin. http: //www.moe.gov.cn/jyb_sjzl/sjzl_fztjgb/202303/t20230323_1051205.html
[2] Smith, J., & White, P. (2020). High school GPA and first-year university performance as predictors of graduation. Research in Higher Education, 61(2), 123-145.
[3] Li, Y., & Carroll, D. (2022). Need-based grants and college persistence: Evidence from a state aid program. Economics of Education Review, 89, 102217.
[4] Biswas, A. A., Majumder, A., Mia, M. J., & Nowrin, I. (2019). Predicting the enrollment and dropout of students in the post-graduation degree using machine learning classifier. International Journal of Innovative Technology and Exploring Engineering, 8(11), 3083-3088.
[5] Mello, Z. R., & Wolfson, A. R. (2011). The role of sleep and circadian rhythms in adolescent health and academic performance. Progress in Brain Research, 191, 289-305.
[6] Romer, D., & Duckworth, A. L. (2015). Academic performance and self-control in adolescents: The role of choice. Journal of Research on Adolescence, 25(2), 313-328.
[7] Dynarski, S. M. (2008). The financial aid cliff: How meeting need could reduce college dropout. Journal of Human Resources, 43(2), 343-370.

DOI: 10.54254/2755-2721/2025.BJ27254

[8] Ferreyra, M. M., Garriga, C., & Martin, J. D. (2015). Raising college access and completion: How much can free college help? (World Bank Policy Research Working Paper No. 7357). World Bank.
[9] Li, X., Zhang, G., & Yan, L. (2021). Predicting student dropout with educational data mining: A systematic literature review. Computers & Education, 174, 104297.
[10] Bowyer, K. W., & Chawla, N. V. (2023). SMOTE: Past, present, and future of synthetic oversampling. IEEE Intelligent Systems, 38(1), 88-95.
[11] Wang, S., Chen, H., & Zhang, N. (2022). Robustness analysis of machine learning models for student dropout prediction. Computers & Education, 186, 104583.
[12] Bawa, P. (2021). Retention in higher education: Using predictive analytics to support student success. Journal of Higher Education Policy and Management, 43(4), 353-367.
[13] Zhao, C.-M. (2022). From reactive to proactive: Toward a predictive analytics framework for student success in Chinese universities. Studies in Higher Education, 47(9), 1896-1912.