

Analysis of Motivating Factors for University Students' Study Abroad Decisions and Research on a Multimodal Classification Prediction Model

Lianrui Chu, Wanjun Wang*, Hui Deng, Qiuxue Ouyang, Conghan Zhong, Yan Yan

College of Computer Engineering, Qingdao City University, Qingdao, China, 266106

*Corresponding author: wanjun.wang@qdc.edu.cn

Abstract: Addressing the challenges of multidimensional factors influencing university students' study abroad decisions and the extreme imbalance in sample distribution, this study constructs a predictive framework integrating feature decoupling, data balancing, and multi-algorithm comparison. The study first performs refined preprocessing on 1,645 multidimensional student feature data points, identifying core positive-correlated drivers of study abroad such as annual household income, family history of overseas residence, and CET-6 scores. To address the severe imbalance where negative samples vastly outnumber positive ones, SMOTE oversampling is introduced to reconstruct the feature space, significantly enhancing the model's accuracy in capturing positive study abroad samples. By integrating K-Means clustering with t-SNE dimensionality reduction, the study revealed the latent structural distribution within the student population. At the algorithmic level, it compared the efficacy of logistic regression, support vector machines, and random forest classifiers. Experiments demonstrated that random forest exhibits optimal generalization capabilities for handling such nonlinear complex relationships, achieving prediction accuracies of 0.8731 for intention prediction and 0.9042 for behavior prediction. This study provides a scientific quantitative basis for international talent cultivation and targeted resource allocation in higher education institutions.

Keywords: Random Forest; SMOTE Oversampling; Study Abroad Intention Prediction

1. Introduction

Against the backdrop of global higher education internationalization, deeply analyzing the mechanisms influencing university students' study abroad decisions holds significant strategic importance for optimizing teaching resource allocation and advancing institutional internationalization[1-2]. However, study abroad decisions represent a complex nonlinear process driven by family economic background, individual academic aptitude, and social capital. In practical educational research, the proportion of students with explicit study-abroad intentions or who ultimately pursue such opportunities remains extremely low. This scarcity often traps statistical and machine learning prediction models into overemphasizing broad sample categories, resulting in poor model performance for identifying critical positive samples. Previous studies predominantly relied on basic descriptive statistics or simple linear regression, struggling to effectively address multicollinearity among high-dimensional features and biases arising from severe sample imbalance. The innovation of this section lies in pioneering the application of the SMOTE oversampling algorithm to correct the distribution of study abroad datasets. It proposes a hybrid research paradigm integrating unsupervised clustering detection with supervised learning classification, aiming to enhance predictive sensitivity for minority samples through algorithmic optimization. The general research approach is as follows: First, the original dataset undergoes multiple rounds of cleaning and normalization, with Pearson correlation coefficients analyzing the contribution of 24 feature dimensions; Second, the student population is clustered using the K-Means algorithm, supplemented by t-SNE for high-dimensional data visualization[3-4]. Finally, by constructing and training logistic regression, support vector machine, and random forest models, performance metrics before and after oversampling are compared within a cross-validation framework to determine the optimal prediction scheme [5].

Data analysis indicates that family history of overseas residence and annual household income are the most strongly correlated positive factors, reflecting the dominant role of family social capital and

economic foundation in study abroad decisions. Additionally, students' participation in teaching and research projects and their demonstrated initiative are significantly positively correlated with study abroad intentions. By applying maximum-minimum normalization and dummy variable treatment to the data, the study eliminated differences in measurement scales. In classification algorithm experiments, the pre-oversampling model exhibited near-zero recall for positive samples. After applying the SMOTE algorithm, the random forest classifier achieved an accuracy of 0.8731 in intention prediction, demonstrating exceptional robustness in precision and F1 scores[6-7]. This validates the algorithm's effectiveness in handling complex imbalanced educational datasets[8].

Attribution and Robust Prediction Modeling of Student Actual Study Abroad Behavior Across Spatiotemporal Dimensions Building upon the intent prediction, this study further modeled students' actual decision to study abroad. Correlation analysis revealed that beyond family economic factors, characteristics such as college entrance exam foreign language scores, CET-6 scores, and gender significantly influence actual study abroad behavior. Notably, male students exhibited higher correlation with the likelihood of studying abroad within the sample. After training on 1,621 valid samples and SMOTE data augmentation, the Random Forest model achieved a prediction accuracy of 0.9042 for actual study abroad behavior, with cross-validation average scores consistently high at 0.9069. Compared to logistic regression (0.8065) and support vector machines (0.8084), the random forest demonstrated overwhelming superiority in balanced prediction performance across positive and negative samples, providing precise decision-making references for applied universities to assess students' final destinations[9-10].

2. Research Implementation Process

2.1 Data Analysis and Visualization

A total of 1645 data entries with 24 dimensional features were collected for this study, and 1621 valid entries remained after removing invalid data. Invalid data included cases such as CET-4 scores exceeding 700 points and parents' ages below 30 years old.

This study involved two tasks: predicting students' intention to study abroad and predicting whether students would ultimately go abroad. Students' intention to study abroad was obtained through questionnaire survey data, while the label of whether students ultimately went abroad was manually assigned by the research team. Based on data analysis, the distribution of positive and negative samples for students' intention to study abroad and their actual overseas study behavior is shown as follows:

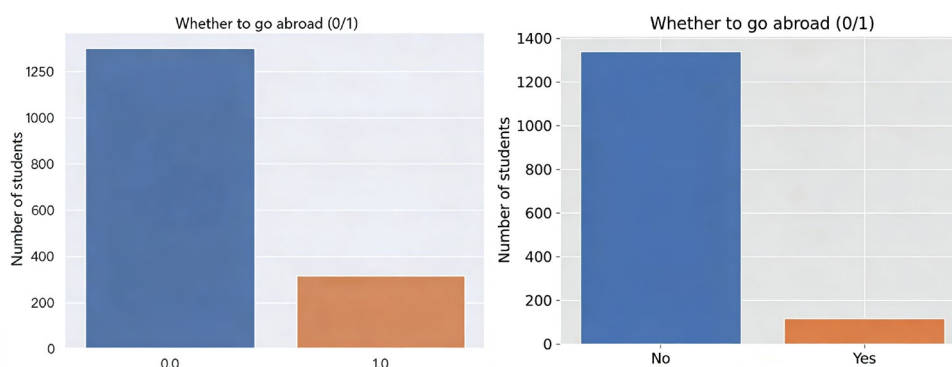


Figure 1 Distribution of Positive and Negative Samples

As can be seen from the figure 1, the distribution of positive and negative samples for both overseas study behavior (0/1) and intention to study abroad is highly imbalanced, with negative samples being approximately 6 to 10 times the number of positive samples. In subsequent research, sample balancing processing of the data is required, usually through oversampling the minority class or undersampling the majority class. Given the relatively small size of the collected data, the oversampling method was adopted in this study.

A total of 24 features were collected, and it was necessary to analyze the relationship between each feature and the two target variables (overseas study behavior and intention to study abroad). After numerical processing of the features, the correlation between each feature and the two target variables was obtained.

The specific correlation results are shown in Figure 2:

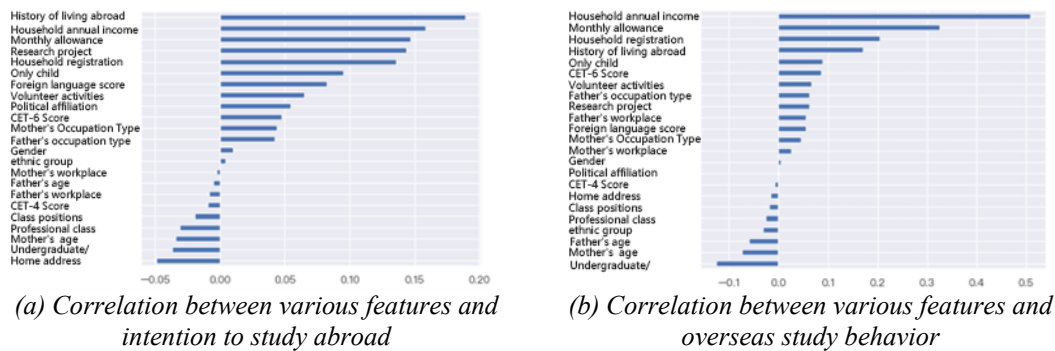


Figure 2(a)(b) Correlation between Various Features and Overseas Study Behavior as well as Intention to Study Abroad

Figure 2 presents the correlation between various features and the two target variables (overseas study behavior and intention to study abroad).

From the positively correlated features shown in Figure 2(a), the feature with the strongest correlation with intention to study abroad is whether family members have overseas living experience, which is consistent with common sense, as family members' overseas experience exerts a significant influence on other family members. The second is family annual income, which also aligns with common sense—studying abroad involves substantial expenses, which only families with relatively high income can afford. The third is students' monthly living expenses, for the same reason as family annual income. Participation in teaching and research projects reflects students' subjective initiative and creativity. The research is conducted in an application-oriented university, which emphasizes students' practical and professional technical abilities; students who participate in research projects demonstrate strong subjective initiative and are thus more likely to be willing to explore studying abroad. Students with non-agricultural household registration are more inclined to study abroad, as their living environment, received education and surrounding influences may all shape their willingness to go abroad. Only children have a stronger intention to study abroad, because families with a single child usually have lower financial burdens and more energy to support their children's overseas study expenses. Participation in volunteer service activities is similar to participation in teaching and research projects in its correlation with study abroad intention. Higher CET-6 scores are associated with a stronger intention to study abroad, while ethnicity has little correlation with the intention to study abroad.

From the negatively correlated features shown in Figure 2(a), lower educational attainment is associated with a higher tendency to study abroad, younger ages of fathers and mothers correspond to a stronger intention to study abroad, and surprisingly, lower CET-4 scores are linked to a higher intention to study abroad.

Figure 2(b) shows the correlation between various features and overseas study behavior (0/1), with the note that the label for overseas study behavior (0/1) was manually assigned by the research team. As can be seen from the figure, family annual income, students' monthly living expenses, household registration type, whether family members have overseas living experience, whether the student is an only child, CET-6 scores, participation in volunteer service activities, participation in teaching and research projects, college entrance examination foreign language scores and gender are positively correlated with overseas study behavior. Political status has no correlation with overseas study behavior, while CET-4 scores, ethnicity, fathers' ages, mothers' ages and educational attainment are negatively correlated with overseas study behavior. In other words, students with higher family annual income, higher monthly living expenses, non-agricultural household registration, family members with overseas living experience, being an only child, higher CET-6 scores, experience in volunteer service activities, participation in teaching and research projects and higher college entrance examination foreign language scores have a greater probability of studying abroad. In terms of gender correlation, male students are more likely to study abroad. On the contrary, students with lower CET-4 scores, ethnic minority status, younger parents and lower educational attainment have a greater probability of studying abroad. Except for the CET-4 score dimension, all other correlations are consistent with common sense.

2.2 Data Preprocessing

2.2.1 Data Cleaning

To ensure data quality, the dataset was cleaned in multiple aspects. First, valueless variables in the dataset were eliminated, such as invalid data in the last column. Meanwhile, samples with missing values were deleted to ensure data integrity. For records marked as "None" in the CET-4 and CET-6 score columns, the values were converted to 0; for non-numeric data in the father's age and mother's age columns, the mean values of the respective columns were used for imputation. In addition, abnormal records with CET-6 scores exceeding 710 were filtered out to avoid interference with subsequent model training.

2.2.2 Data Transformation

Since machine learning models can only process numerical data, and the dataset contains a large number of categorical variables described in text (such as gender, major and class, ethnicity, etc.) that cannot be directly accepted by the models, a function was established to perform one-hot encoding on these categorical variables, converting them into numerical variables for the models to understand and process.

2.2.3 Data Normalization

To eliminate the dimensional differences between different features, min-max normalization was applied to CET-4 and CET-6 scores, scaling their value ranges uniformly to 0-1.

2.3 Research Design and Model Construction

2.3.1 Clustering Analysis for Data Distribution

As an unsupervised learning method, cluster analysis aims to divide samples in a dataset into different groups or clusters, such that samples within the same cluster have high similarity and samples between different clusters have significant differences. In the analysis of this study abroad dataset, cluster analysis helps explore the potential structure and patterns of the data, understand the characteristics of different student groups, and thus provide valuable references for subsequent decision-making and analysis.

The K-Means clustering algorithm was selected for this study. As a commonly used distance-based clustering algorithm, it assigns data points to different clusters through an iterative process, with the goal of minimizing the sum of distances from data points to the centroid of their respective clusters.

Cluster analysis was performed on the preprocessed data to ensure data quality and consistency. All columns except the "study abroad" column were used as the input dataset, and the "study abroad" column was set as the target variable. The K-Means clustering algorithm was used to divide the dataset into 4 clusters, the clustering results were saved in a CSV file, and corresponding visualizations were generated (see Figure 3).

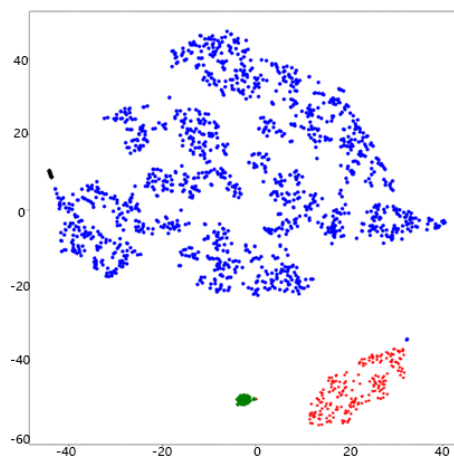


Figure 3 Clustering Results - 4 Clusters

For a more intuitive observation of the clustering results, t-SNE technology was used for dimensionality reduction of the data, reducing the high-dimensional data to two dimensions for visualization (see Figure 4).

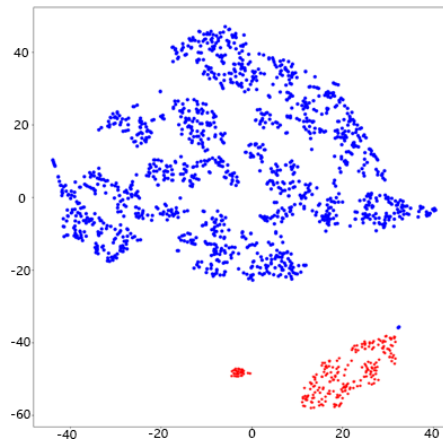


Figure 4 Clustering Results - 2 Clusters

2.3.2 Feature Selection

Feature selection is crucial for model construction. In this study, a series of columns potentially associated with students' intention to study abroad were selected as input features, including "undergraduate/college/junior college to undergraduate", "whether the student is an only child", "household registration type", while relatively unimportant data such as parents' ages were excluded. For example, in terms of family economic status, family annual income and students' monthly living expenses may affect students' financial ability to study abroad; in terms of academic performance, CET-4 scores, CET-6 scores and college entrance examination foreign language scores can reflect students' English proficiency and learning ability to a certain extent, thereby influencing their intention to study abroad. Reasonable feature selection not only reduces the time and complexity of model training but also avoids model overfitting caused by excessive irrelevant features, improving the prediction accuracy and generalization ability of the model.

2.3.3 Model Selection

Three models were selected for modeling in this study: Logistic Regression, Support Vector Machine (SVM) and Random Forest Classifier, for the following specific reasons:

Logistic Regression: a linear classification model that predicts sample categories by fitting a linear equation. It features relatively simple calculation processes and strong interpretability, which can clearly show the direction and degree of influence of each feature on the prediction results. However, it has weak fitting ability for nonlinear data and may lead to poor prediction performance if there are complex nonlinear relationships in the data.

Support Vector Machine: realizes classification tasks by finding the optimal classification hyperplane and performs excellently in processing high-dimensional and nonlinear data. Nevertheless, this model has high computational complexity and is sensitive to parameter selection; inappropriate parameter settings may affect model performance.

Random Forest Classifier: an ensemble learning model based on decision trees, which constructs multiple decision trees and synthesizes their prediction results. It has good generalization ability and anti-overfitting ability, but its interpretability is relatively poor due to the complex model structure, making it difficult to intuitively understand the decision-making process of the model.

3. Research Results

3.1 Model Evaluation

3.1.1 Prediction of Intention to Study Abroad

In the initial stage of model training, the proportion of positive samples (students with the intention to study abroad) in the dataset was relatively low. This sample imbalance may lead to poor accuracy of the model in predicting positive samples. To effectively improve this situation, the SMOTE method was used in this study to perform oversampling on positive samples. This method generates new sample points in the feature space of positive samples to increase the number of positive samples, so as to balance the ratio of positive and negative samples in the dataset. After rigorous model training and comprehensive

evaluation, the following results were obtained (table 1):

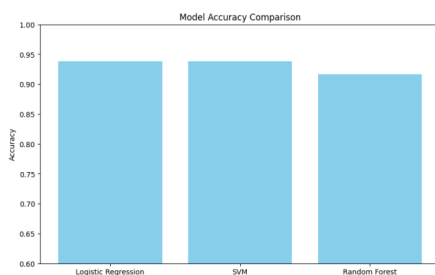
Table 1 Classification Report before Oversampling

Model	Accuracy	Classification Report	Cross-validation Average Score
Logistic Regression	0.9352	The number of positive test samples is 304.0, with precision, recall and F1-score all 0; The number of negative test samples is 304.0, with precision 0.9381, recall 0.9967 and F1-score 0.9665	0.9245
Support Vector Machine (SVM)	0.9383	The number of positive test samples is 304.0, with precision, recall and F1-score all 0; The number of negative test samples is 304.0, with precision 0.9383, recall 1.0 and F1-score 0.9682	0.9258
Random Forest	0.9352	The number of positive test samples is 304.0, with precision, recall and F1-score all 0; The number of negative test samples is 304.0, with precision 0.9381, recall 0.9967 and F1-score 0.9665	0.9227

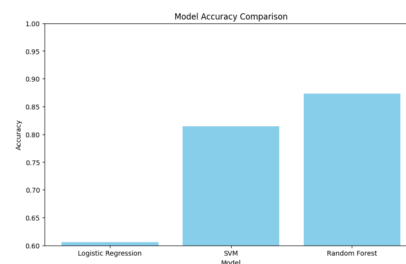
By comparing the accuracy of each model before and after oversampling, it can be clearly seen that oversampling has significantly changed the data distribution, thereby exerting a positive impact on the model results and making the prediction results more reliable and convincing. The performance of each model after oversampling is shown as follows (table 2):

Table 2 Classification Report after Oversampling

Model	Accuracy	Classification Report	Cross-validation Average Score
Logistic Regression	0.6060	The number of positive test samples is 299.0, with precision 0.6113, recall 0.5786 and F1-score 0.5945; The number of negative test samples is 300.0, with precision 0.6013, recall 0.6333 and F1-score 0.6169	0.6106
Support Vector Machine (SVM)	0.8147	The number of positive test samples is 299.0, with precision 0.7848, recall 0.8662 and F1-score 0.8235; The number of negative test samples is 300.0, with precision 0.8513, recall 0.7633 and F1-score 0.8049	0.8173
Random Forest	0.8731	The number of positive test samples is 299.0, with precision 0.8348, recall 0.9298 and F1-score 0.8797; The number of negative test samples is 300.0, with precision 0.9211, recall 0.8167 and F1-score 0.8657	0.8791



(a) Comparison of model accuracy before oversampling



(b) Comparison of model accuracy after oversampling

Figure 5 Comparison of Model Accuracy

In Figure 5(a), Logistic Regression and SVM models have similar accuracy, both around 0.93-0.94, while the Random Forest model is slightly lower, around 0.91-0.92; in Figure 5(b), the Random Forest model has the highest accuracy at approximately 0.87, followed by the SVM model at approximately 0.81, and the Logistic Regression model is the lowest, just over 0.60.

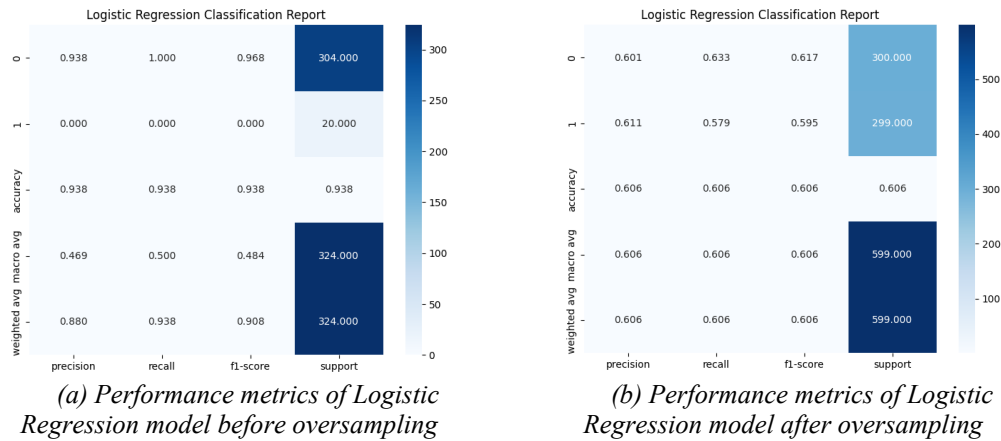


Figure 6(a)(b) Performance Metrics of Logistic Regression Model

The overall accuracy of Figure 6(a) is 0.938, with macro-average precision, recall and F1-score of 0.469, 0.500 and 0.484 respectively; the overall accuracy of Figure 6(b) is 0.606, with macro-average precision, recall and F1-score all 0.606.

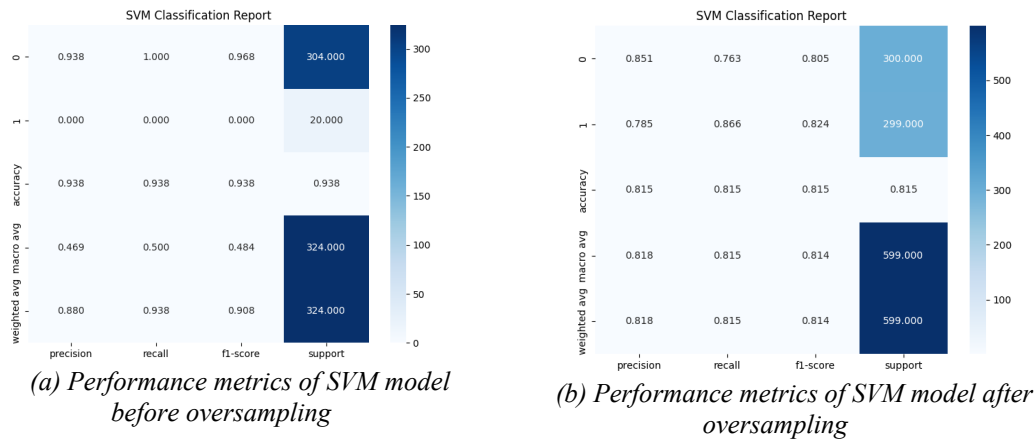


Figure 7(a)(b) Performance Metrics of Support Vector Machine Model

The overall accuracy of Figure 7(a) is 0.938, with macro-average precision, recall and F1-score of 0.469, 0.500 and 0.484 respectively; the overall accuracy of Figure 7(b) is 0.815, with macro-average precision, recall and F1-score of 0.818, 0.815 and 0.814 respectively.



Figure 8(a)(b) Performance Metrics of Random Forest Model

The overall accuracy of Figure 8(a) is 0.917, with macro-average precision, recall and F1-score of 0.562, 0.535 and 0.543 respectively; the overall accuracy of Figure 8(b) is 0.873, with macro-average precision, recall and F1-score all 0.873.

3.1.2 Prediction of Overseas Study Behavior

Similar to the model for predicting the intention to study abroad, the positive and negative samples for overseas study behavior are highly imbalanced, with 306 positive samples and 1301 negative samples. Therefore, the SMOTE method was also applied to perform oversampling on positive samples for the prediction of overseas study behavior. This method generates new sample points in the feature space of positive samples to increase the number of positive samples, so as to balance the ratio of positive and negative samples in the dataset. After rigorous model training and comprehensive evaluation, the following results were obtained (table 3):

Table 3 Classification Report on Overseas Study Behavior Prediction (Before Oversampling)

Model	Accuracy	Classification Report	Cross-validation Average Score
Logistic Regression	0.8430	The number of positive test samples is 64, with precision 0.6757, recall 0.3906 and F1-score 0.4950; The number of negative test samples is 261, with precision 0.8646, recall 0.9540 and F1-score 0.9071	0.8383
Support Vector Machine (SVM)	0.8184	The number of positive test samples is 64, with precision 0.8571, recall 0.0937 and F1-score 0.1690; The number of negative test samples is 261, with precision 0.8176, recall 0.9961 and F1-score 0.8981	0.8118
Random Forest	0.9352	The number of positive test samples is 304.0, with precision, recall and F1-score all 0; The number of negative test samples is 304.0, with precision 0.9381, recall 0.9967 and F1-score 0.9665	0.9227

By comparing the accuracy of each model before and after oversampling, it can be clearly seen that oversampling has significantly changed the data distribution, thereby exerting a positive impact on the model results and making the prediction results more reliable and convincing. The performance of each model after oversampling is shown as follows (table 4):

Table 4 Classification Report on Overseas Study Behavior Prediction (After Oversampling)

Model	Accuracy	Classification Report	Cross-validation Average Score
Logistic Regression	0.8065	The number of positive test samples is 252, with precision 0.7807, recall 0.8333 and F1-score 0.8061; The number of negative test samples is 270, with precision 0.8340, recall 0.7815 and F1-score 0.8068	0.8022
Support Vector Machine (SVM)	0.8084	The number of positive test samples is 252, with precision 0.7695, recall 0.8611 and F1-score 0.8127; The number of negative test samples is 270, with precision 0.8541, recall 0.7593 and F1-score 0.8039	0.8123
Random Forest	0.9042	The number of positive test samples is 252, with precision 0.8797, recall 0.9286 and F1-score 0.9035; The number of negative test samples is 270, with precision 0.9297, recall 0.8814 and F1-score 0.9049	0.9069

3.2 Analysis of Research Results

Among the three models, due to the influence of sample imbalance before oversampling, the performance metrics of class 0 were high, while those of class 1 were extremely low; after oversampling, the performance metrics of the two classes were relatively balanced, making the analysis results more universal.

From the data analysis after oversampling, the Random Forest Classifier is the most suitable model selection for the prediction of students' intention to study abroad based on the comprehensive evaluation of various metrics.

First, in terms of accuracy comparison among the three models, the Random Forest Classifier reached 0.8731 after oversampling, which is significantly higher than 0.6060 of Logistic Regression and 0.8147 of SVM. Second, in terms of precision, recall and F1-score, the Random Forest Classifier performed excellently in predicting both positive and negative samples. For positive samples, its precision was 0.8348, recall was 0.9298 and F1-score was 0.8797; for negative samples, the precision was 0.9211, recall was 0.8167 and F1-score was 0.8657. Third, in terms of cross-validation average score, the Random Forest Classifier reached 0.8791, which is higher than 0.6106 of Logistic Regression and 0.8173 of SVM.

Based on the comprehensive results, the Random Forest Classifier can be adopted for the model of predicting students' intention to study abroad. It shows obvious advantages in accuracy, prediction performance of positive and negative samples as well as generalization ability, which can provide more reliable and effective results for the prediction of students' intention to study abroad and contribute to relevant research and decision-making analysis.

4. Conclusions

This study systematically identified core factors driving university students' study abroad decisions by integrating data mining, sample balancing, and machine learning algorithms, constructing a high-precision model for predicting study abroad intentions and behaviors. The research confirms the superiority of the Random Forest classifier combined with SMOTE oversampling technology in handling imbalanced educational datasets, achieving accuracy and robustness that meet practical application standards. However, this study has certain limitations. First, the relatively small volume of collected raw data and the concentrated feature distribution may have somewhat impacted the model's generalization capability across broader domains. Second, the atypical negative correlations observed in certain features like CET-4 scores suggest potential data noise when reflecting specific student cohorts. Future research should focus on expanding sample sizes and exploring advanced frameworks such as deep learning or ensemble learning to capture deeper feature interactions. This approach will provide more intelligent and precise guidance for student career planning and international education management in higher education institutions.

Acknowledgements

This research was completed under the support of the project “Research and Application on Precise Mining of Potential Undergraduate Study-Abroad Groups Based on Artificial Intelligence” (Project No. 2024004A). Project Source: 2024 Qingdao City University University-Level Teaching Research Project

References

- [1] Xu Haowen, Sun Jianrong, Chen Qian. *Empirical Analysis of Factors Influencing International Students' Decision to Study in China: Strategies for Enhancing the Attractiveness of the “Study in China” Brand* [J]. *Comparative Education Research*, 2025, 47(07): 23-35.
- [2] Li Xianjiang, Zhang Ziyi, Li Yan. *A Study on the Driving Factors of Returning Overseas Students' Intentions for Science and Technology Entrepreneurship: An S-O-R Theory Perspective* [J]. *China Science and Technology Forum*, 2025, (06): 160-166.
- [3] Qin Huimin, Xiatiguli Shakier. *Manifestations and Influencing Factors of Cross-Cultural Guidance Behaviors among Advisors: A Qualitative Study of Advisors to International Postgraduate Students at a “Double First-Class” University* [J]. *Degree and Postgraduate Education*, 2025, (03): 24-31.
- [4] Li Junru, Li Xueyong. *Political Guidance for Returning Overseas Talents in Higher Education: Contemporary Value, Problem Analysis, and Practical Approaches* [J]. *Science of University Education*, 2025, (02): 68-77.
- [5] Li Fengliang, Liu Manqing. *Research on the Impact Mechanism of Graduate Education for International Students in China on Industrial Structure Upgrading* [J]. *Research on Graduate Education*, 2024, (06):13-23.
- [6] Li Tao, Zhang Jun, Cai Juan. *The Impact of Education for International Students in China on Enhancing China's National Image* [J]. *Research in Higher Education*, 2024, 45(10):24-36.
- [7] Zhang Fengjuan, Zhang Zihan. *International Hotspots and Development Trends in Second Language Learning Research in Study Abroad Contexts* [J]. *Foreign Language World*, 2024, (05): 82-89.
- [8] An Yaluen, Huang Shihan, Liu Baocun. *Japan's Experience and Implications for Alleviating*

Outbound Student Mobility Challenges [J]. Shanghai Education Research, 2024, (10): 62-70.

[9] Shen Wenqin. *Transformation and Response: Research on Independent Cultivation of Elite Innovative Talents among Postgraduates amid New Trends in Overseas Study [J]. Research on Chinese Higher Education, 2024, (04): 85-92.*

[10] Shi Yue, Tian Yu, He Dexu. *Overseas Education Background of Independent Directors and Corporate ESG Performance [J]. Research on Financial Issues, 2024, (02): 76-89.*