**Application result prediction**

Contents

[1. Abstract 1](#_Toc521603631)

[2. Introduction 1](#_Toc521603632)

[3. Theoretical Methods 2](#_Toc521603633)

[3.1 Random forest 2](#_Toc521603634)

[3.2 GBDT 3](#_Toc521603635)

[3.3 XGBoost 3](#_Toc521603636)

[3.4 Stacking 3](#_Toc521603637)

[4. Results and Discussions 4](#_Toc521603638)

[4.1. Data Analysis 4](#_Toc521603639)

[4.2. Model Performance 4](#_Toc521603640)

[5. Conclusion 5](#_Toc521603641)

[6. Reference 5](#_Toc521603642)

# Abstract

Along with the upgrading of people's living standard in the 21st century, more and more people choose to study abroad. However, it is hard for students to predict whether they can be accepted, and too many application materials are a burden for universities, so it is desirable to have an artificial intelligence system, which can help the students to choose school scientifically and help schools to select students efficiently. In this work, we firstly transform the four-classification task into a binary classification task as the confusion matrix shows that models are difficult to classify class 0 and class 1. After getting the binary model, we have also trained the three-classification model and the four-classification model. Last, we ensemble these model’s outputs as the result. Result suggests that this ensemble model has better performance than the individual four-classification model and the average accuracy on the 10 folds cross-validated dataset is 89.3%. Based on this result, we can conclude that the artificial intelligence plays an important role in the field of educational technology.

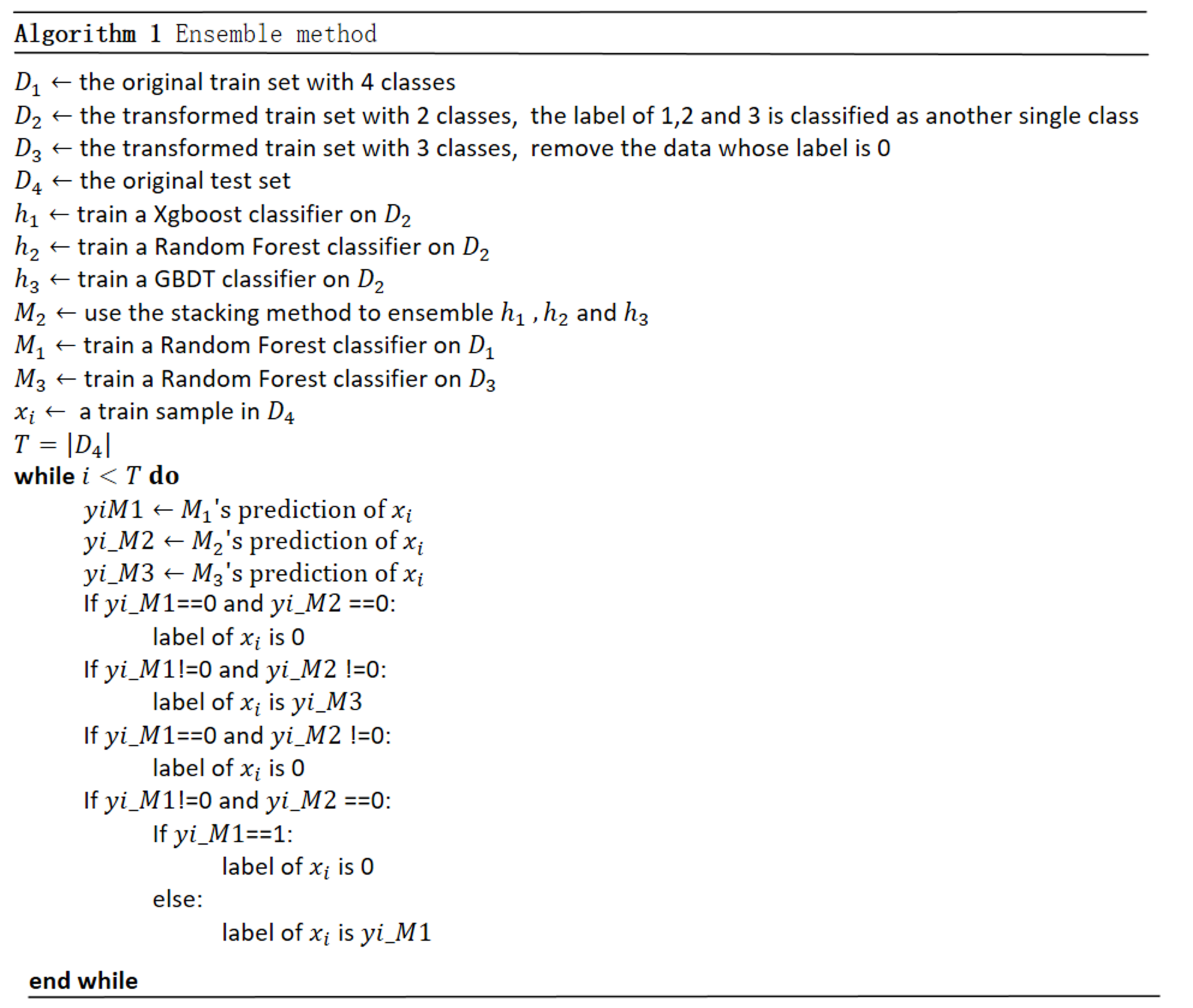
**Key Words**: Artificial intelligence, Ensemble, Educational technology

# Introduction

In March 2016, the Chinese Ministry of Education published statistical data, which shows that from 1978 to the end of 2015, China's cumulative number of overseas students was 4.0421 million, with an average annual growth rate of 19.06%. In 2015, a total of 520 thousand Chinese students went abroad to study, which has increased nearly 14% over the previous year [1]. Although the trend of younger age has gradually appeared in recent years, college graduates still account for a large proportion of the upsurge of studying abroad [2].

As we know, the success of the application for studying abroad is not just decided by a single factor, generally, each condition of the students will be taken into consideration. As a result, for the university, a large amount of manpower and material resources will be consumed to fairly select the students they need. For students, it's hard for them to predict whether they can be admitted, it’s a waste of time and money if the application fails. Besides, the traditional method determines the application completely by people, thus a lot of subjective judgments will be introduced. So, if we train a machine learning system with the data of the students who have been admitted before, the application procedure will be more effective and objective.

For the reasons above, we have developed a method to accurately predict the result of a student’s application. This method has the following several steps: First, as the confusion matrix shows that the four-classification model is difficult to classify by label 0 and label 1, we transform the four-class classification task into binary classification task, then we use the XGBoost, GBDT, and Random forest as base model to get the ensemble model named M1. Second, we have also used the Random Forest algorithm to train the four-classification model named M2 and three-classification model named M3. The following Pseudocode shows the full procedure in details.



# Theoretical Methods

All the calculations and modeling were performed on the YouFang AI Cloud Platform. First, we used Numpy and Scipy to do scientific calculations, Pandas is applied for data analysis, Matplotlib and Seaborn are used for data visualization. Besides, we used the XGBoost and Scikit-learn package to construct the XGBoost, Random forest and GBDT model. Last, the Scikit-learn package is also used for measuring the performance of the model.

# 3.1 Random forest

The random forest is an ensemble model [3] consisting of multiple decision trees. In general, the decision tree is a Cart tree. It combines the results of multiple decision trees and solves the shortcoming of a single decision tree that a single model is easy to overfit the training set [4].

Assuming that our training data set is, the feature dimension is , the number of randomly selected features is , the decision tree learning algorithm is and the number of decision trees used is . This method has the following several steps:

1. Bootstrap sampling is used to obtain the size sampling data set from the training dataset .

2. features are randomly selected from features. Based on these selected features and , a decision tree is obtained by training the decision tree model. [5]

3. For the classification problem, we export the final ensemble model as follows:

(1)

# 3.2 GBDT

On contrary to the Random forest, who aims at reducing model’s variance, GBDT is a boosting method combining weak "learners" into a single strong learner in an iterative fashion, such a procedure will efficiently reduce the bias of a model [4].

Assuming that the training set is (, ), the number of iterations is and the loss function is . This method has the following several steps:

1. Initialize a model named .

2. Compute the so-called pseudo-residuals:

(2)

3. Fit the t-th decision tree with :

(3)

4. Use line search to find the best step:

(4)

5. Set , and update the model .

6.Loop through the process above times until we get the final model .

# 3.3 XGBoost

Compared with GBDT, the second derivative of the loss function is used in XGBoost. Besides, the regularization term is also added to the loss function, which helps to control the complexity of the model and avoid the over-fitting [6].

Assuming that is the number of the training, is the prediction of the former t-1 models, the regularization term of model is, the constant term is and the loss function is.

The first partial derivatives of is as followed:

(5)

The second partial derivatives of is as followed:

(6)

Then the final objective function has been transformed into:

(7)

# 3.4 Stacking

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone [7], and stacking is a model ensembling technique used to combine information from multiple predictive models to generate a new model.

This method has the following several steps:

1. Split the training data into k folds.
2. Train each of the L base algorithms on the training folds.
3. Collect the cross-validated predicted values from each of the L algorithms.
4. Train the meta-learning algorithm on the predicted values.

# Results and Discussions

## Data Analysis

The train data we get has 900 samples, and the number of features is 15. The column name of label is rank and this column has four different values. In order to know whether the class is balanced, we first analyzed the label, as shown in Fig.1.

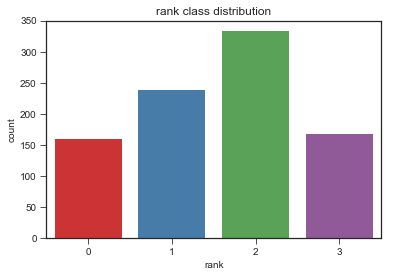
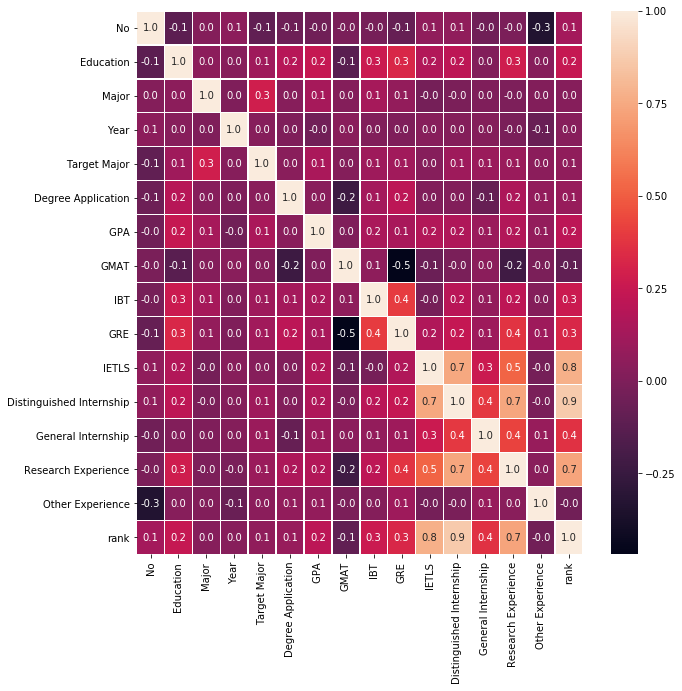
 

Figure 1. Rank class histogram Figure 2. Correlation between each column

From the figure 1 above, we find that the class is relatively balanced so there is no need to use other methods for unbalanced data classification. Besides, we also used the Seaborn to analyze the correlation between each feature and the label, as shown in Fig.2. Through the heat map above, we conclude that IETLS, distinguished Internship and research experience all have great influence on the rank. We also find a student’s distinguished Internship is closely related to the IETLS and research experience. All these discoveries will lay a solid foundation for the latter feature engineering.

## Model Performance

Feature engineering has the greatest impact on model performance, thus we spend a lot of time analyzing the data and doing feature engineering. For the features degree application and distinguished internship, we classify them as categorical features and use the one-hot encoding. For the feature research experience, we add a new column to judge if it is greater than 7. Besides, for the numerical feature IBT, we add a new column to judge if it is 0.

Due to the limited generalization ability of a single model, we use an ensemble method to train the binary classification model. After training the binary model, we also trained a three-classification model and the four-classification model and used these three models to get the final output. The final confusion matrix of the cross-validated training set is shown in Fig.3.

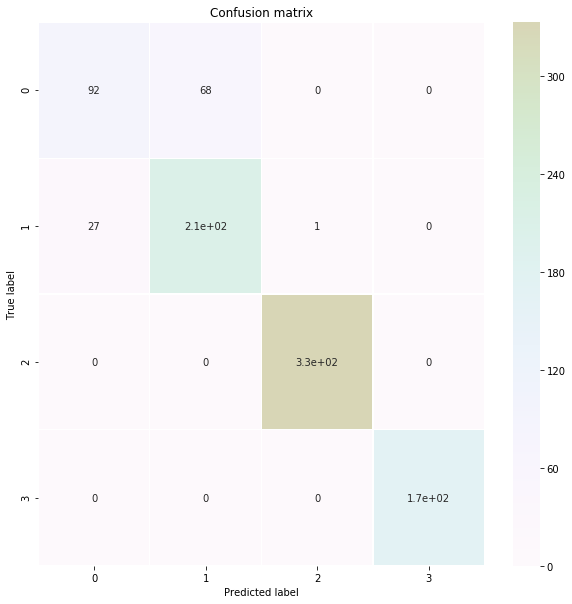


Figure 3. The final model’s confusion matrix on cross-validated data

Figure 3 illustrates that our model has 89.33% accuracy on the 10 folds cross-validated training set. Especially, the result also shows that we can easily classify label 2 and label 3 as only one sample is misclassified.

# Conclusion

Based on the average accuracy on the cross-validated dataset, we can conclude that our ensemble model has a great performance in classifying the application results, especially for rank 2 and rank 3. On the one hand, equipped with such an AI system, the university can examine the applications automatically, thus it can save a large amount of manpower and material cost; on the other hand, based on their conditions, students can scientifically predict the results of their application, which can reduce their cost of trial-and-error. As a result, artificial intelligence can play an important role in the field of educational technology.

# Reference

[1]*《2016中国出国留学发展趋势报告》*.教育部.

[2] 陆根书, 田美, 黎万红.大学生出国留学意愿的影响因素分析 [J].复旦教育论坛,2014(5).

[3] Ho, Tin Kam. *Random Decision Forests*, *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14–16 August 1995. pp. 278–282.

[4] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. *The Elements of Statistical Learning* (2nd ed.). *Springer*. ISBN 0-387-95284-5.

[5] 李航，*统计学习方法*，清华大学出版社，2012.

[6] Chen, Tianqi; Guestrin, Carlos (2016). "XGBoost: A Scalable Tree Boosting System". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[7] Polikar, R. (2006). "Ensemble based systems in decision making". *IEEE Circuits and Systems Magazine*. 6 (3): 21–45. doi:10.1109/MCAS.2006.1688199.