Research on the Factors Influencing the Performance of the Art College Entrance Examination Based on Artificial Intelligence Technology

Mengying Ma

[1]*Department of Cinematography, Beijing Film Academy, Beijing, China*
mmyenjoy@163.com

Abstract
The application of artificial intelligence technologies has influenced many aspects of education, including the learning style of students, the teaching of education faculty, and the administration of colleges. However, the research focusing on the art college entrance examination in China is limited, and the research methods are somehow outdated. This article will introduce the state-of-the-art classifying methods for modeling the imbalanced data sets (Imbalanced-Ensemble) and some popular interpretation tools (Permutation Importance and SHAP) to analyze the details of the influencing factors of the art exam performances and NCEE performances of students from one of the most famous art colleges in China: Beijing Film Academy. Many important supplementary data collected manually from the website also play an important role in modeling and interpretation. The conclusions of this article will help more students to make more wise decisions when applying to schools and will help the policymakers of schools to adjust the details of the art exam for promoting the development of the admission of higher education of art.

Keywords*: the art exam, data science, admission strategy, art colleges, artificial intelligence*

1 INTRODUCTION

The national college entrance examination (NCEE) is one of China's most important official tests. Besides the basic subjects of NCEE, liberal art, and science, the art exam is also an important part of NCEE for those people willing to do the jobs related to art. In recent years, with the rapid development of China's economy, more and more parents and students choose to take the art exam. There is even a boom in the art exam, which causes a lot of discussions and research. Based on the data from different training providers and local education bureaus, the number of students taking the art exam is not only significant but also a large proportion of the total students taking the NCEE. More and more traditional colleges set up some new majors related to art, besides the key art colleges with a history of decades, offering more opportunities for students. Due to the excellent teachers and better job prospects, there is a lot of pressure to compete for places in key art colleges. Not only the number of key schools is limited, but the number of planned enrollments is also small (Figure1). Thus, it is very meaningful to analyze the data related to the flow of the art exam. On one hand, students can get important guidance for the application, on the other hand, the staff of the admissions office can tune the rules of the details of the art exam or make plans for the propaganda.

This article focuses on one of the key art colleges of China, Beijing Film Academy (BFA), using the application data, the art exam related data, and final enrollment data in 3 years (2016, 2017, 2018). More than 100,000 data items are used in this paper. The main objectives of this work can be summarized as follows:

(1) Collecting the relevant meaningful data and using state-of-the-art intelligent algorithms (Imbalanced-Ensemble) to predict the students' academic performance on the art exam and the final admission result as accurately as possible and compare their performance using various appropriative metrics.

(2) Using popular interpretable machine learning and tools, permutation importance, and SHAP to understand the decisions of the models, moreover, to find the key factors that affect academic success and how they affect the decision path in detail.

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 590-603, 2023.
https://doi.org/10.2991/978-94-6463-040-4_90

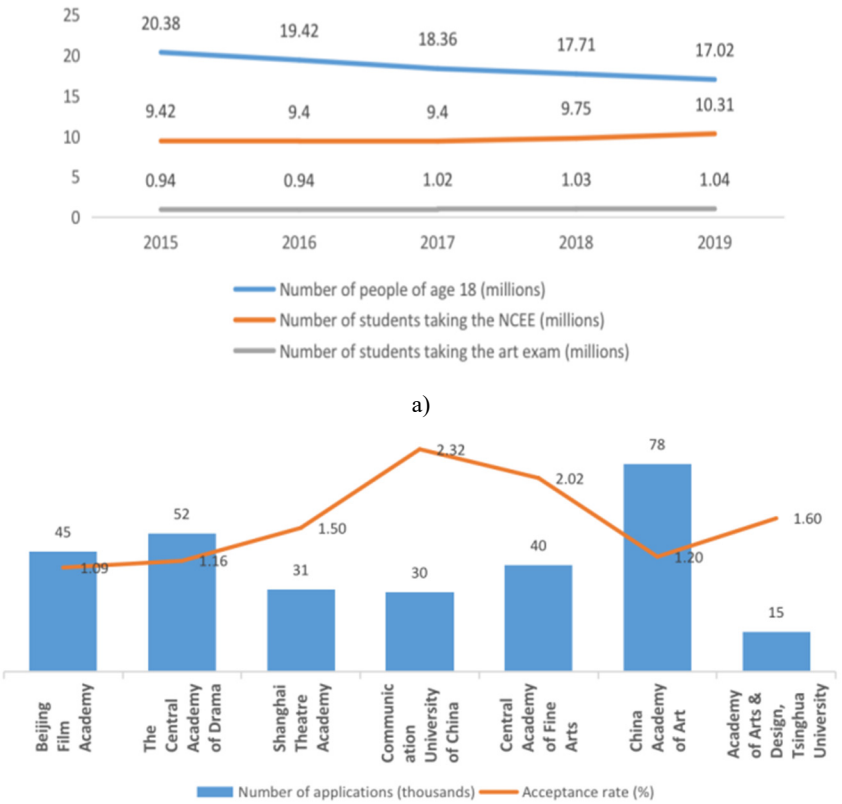


Figure 1: Background information on the art exam. a) Number of students taking the art exam between 2015 and 2019. b) Number of applications and acceptance rate of some key art colleges in 2018 (school names from left to right: Beijing Film Academy, The Central Academy of Drama, Shanghai Theatre Academy, Communication University of China, Central Academy of Fine Arts, China Academy of Art, Academy of Arts & Design of Tsinghua University)

2 DATA SETS AND METHODOLOGY

2.1 The Flow of The Art Exam and The Background of The Admission Work of BFA

To help the readers easier to follow, here we briefly introduce the flow of the art exam in China and some related rules. To get a higher education, most students take the NCEE. Apart from the compulsory subjects (Chinese, Math, and English), most students also need to choose whether to take the liberal art test including History, Politics, and Geography, or the science test including Physics, Chemistry and Biology or choose some interesting subjects in several provinces. The colleges often choose students according to their rankings in different provinces. Students choosing liberal art or science will be ranked separately by score.

Students taking the art exam need to take the art exam before taking the NCEE. There are usually 2 types of the art exam. Those key art colleges such as BFA and The Central Academy of Drama will hold the art exam using their own flow of tests and scoring criteria. Other art colleges will ask students to take the Provincial joint examination of fine arts and take their scores of it to represent their talent. After taking the art exam and the NCEE, both scores will be taken into consideration for colleges to decide whether accept a student.

The enrollment style of BFA is the former pattern. First, the teachers decide the planned enrollment number and the content and criteria of the exam. Students need to take the exam and get a score. Then the students in the top rank will pass the art exam. The number of qualified students is about 3 times the planned enrollment number. Then after they take the NCEE, the school will take their NCEE scores and the art exam score together. According to different criteria, the school then makes the final rankings of students. Usually, the criteria are the different proportions of the NCEE score of the cut-off point of the first batch of universities. Then the top students will be admitted until the planned enrollment number is satisfied.

2.2 Data Description and Feature Engineering

The data used for the experiment are the application data, the art exam related data, and final enrollment data from one of the key art colleges in China, Beijing Film Academy (BFA), in the period 2016 to 2018.

For better modeling and exploiting the use of many basic features, some relevant features have been added. Many of them are proved to be very meaningful in prediction. For example, the original data has a feature called *High School Name*, then more information has been collected to prove the attribute of this feature such as *Whether the student is graduated from art high school* and *Type of high school* of the student. To better describe the exam, the *Difficulty of the art exam* and the *Number of subjects of the art exam* are added. There are also some economy-related features such as *GDP per capita* and *Per capita disposable income*, some culture-related features such as *Number of art organizations*, *Number of art performances,* and education-related features such as *Number of colleges, Number of planned enrollments of all colleges*, *Number of colleges of project 985*, *Number of colleges of project 211*, *Number of Key Art Colleges* added.

After filling the blanks of the data or handling the missing values, the non-numeric features are encoded using one-hot coding, hashing coding, etc. For some newly added features, customized encoding methods according to the practical meaning of values are used. For a better understanding of some unusual features, a necessary explanation is needed. The feature *Major* is encoded using hashing coding method. Since the classes of majors are divided by the precise direction of one big major, the total number of majors is 34, which is relatively large for other encoding methods. The feature *Whether graduated from art high school* is encoded by classifying the type of high school. The schools that are affiliated with the key art colleges and whose name has the keywords: art, music, drama, film, dancing, movie, etc. are identified as art high school. The number of art high schools is 5661, accounting for 5% of the total high school. The feature *Type of high School Graduation* is encoded by using the open data of the Education Bureau of different provinces. The feature *Difficulty of the art exam* is calculated by summing up the difficulty of every subject. The difficulty of the subject which has the standard answer is identified as 1, and the difficulty of other subjects which test students' creative ability or request the interview is identified as 2. The newly added economy-related features, culture-related features, and education-related data are collected through the open data of the National Bureau of Statistics. Relevant research has revealed that the training of a student for taking the art exam is more complex, including more investment in skill training and cultural cultivation [17].

The feature names and descriptions are in Table1.

Table 1: Features Description

Feature Class	Feature Name	Description
Personal Details	Major	The major that the student is applying for
	Year of exam	2016/2017/2018
	Sex	Male or female
	Age	Age of integer
	Politics Status	Normal people, member of Communist Youth League of China, or member of other politics parties
	Nationality	The han nationality or minority nationality
	Type of sources of students	4 classes: new graduates from city (NGC), new graduates from rural area (NGR), other graduates from city (OGC) and other graduates from rural area (OGR)
	Type of the NCEE	Liberal arts, science or other
	Highest Education Qualification	High school diploma or below, junior college, undergraduate degree, master’s degree and doctor’s degree
	Height	Height of the student
	Whether graduated from art high school	31 provinces in mainland China
	Type of high School of Graduation	Demonstrative high school, normal high school, or vocational middle school

	Province where the student took NCEE	
	City where the student took NCEE	Provincial capitals or others
	Household registration location	Province where the student's household registration location is
The art exam related	Score of art exam	Numerical value in the range [0,100]
	Number of art exam subjects	Numerical value in the range [2,4]
	Difficulty of art exam	Numerical value in the range [3,7]
	Number of planned enrollments	The number of planned enrollments of the major in that year
	Rankings in the major	Rankings in the major of application
	Rankings in the major within the same province	Rankings in the major in the province where the student took the NCEE
	Student's preference order for the major	Numerical value in the range [1,3], value 1 for the first choice etc.
Society related	Cut-off point of the first batch of universities	Cut-off point of the first batch of universities in the province where the student took the NCEE
	Cut-off point of art universities	Cut-off point of art universities in the province where the student took the NCEE
	GDP per capita	GDP per capita in the province where the student took the NCEE
	Per capita disposable income	Per capita disposable income in the province where the student took the NCEE
	Number of performances by performing arts organizations	Number of performances by performing arts organizations in the province where the student took the NCEE
	Number of art performances	Number of art performances in the province where the student took the NCEE
	Number of colleges	Number of colleges in the province where the student took the NCEE
	Number of planned enrollments of all colleges	Number of planned enrollments of all colleges in the province where the student took the NCEE
	Number of colleges of project 985	Number of colleges of project 985 in the province where the student took the NCEE
	Number of colleges of project 211	Number of colleges of project 211 in the province where the student took the NCEE

	Number of Key Art Colleges	Number of Key Art Colleges in the province where the student took the NCEE
Target	Whether passed the art exam	0/1
	Admitted	0/1

2.3 Imbalanced Data Sets Analysis Methods

2.3.1 Imbalanced Data Sets Classification Methods

There are 2 experiments in this research. The first one is predicting whether the student is passed the art exam, and the second one is predicting whether the student is finally admitted to the BFA. Because the distribution of the results is heavily imbalanced (0: 1 = 28: 1 for the first experiment, and 0: 1 = 2: 1 for the second one), the imbalanced ensemble (IMBENS) of Python is used [9].

The problem of learning unbiased models from imbalanced data is often referred to as imbalanced learning or long-tailed learning for multi-category learning. Traditional machine learning models assume that the edge distribution P(Y) of the data is approximately uniform, and they are usually designed to optimize the accuracy of classification without considering the difference in sample size across categories.

In the case of imbalanced data, categories with small sample sizes have little effect on classification accuracy, so models that directly optimize classification accuracy will result in poorer prediction results for minority classes. However, minority classes usually carry more important information. We, therefore, hope to use some means to correct the bias introduced to the model by unbalanced data and obtain an unbiased prediction model.

The algorithms for analyzing the imbalanced data used in this paper are as follows.1.Methods using under sampling (discarding some samples from the majority class): SelfPacedEnsembleClassifier (SPEC) [8], BalanceCascadeClassifier (BCC) [7], BalancedRandomForestClassifier (BRFC) [4], EasyEnsembleClassifier (EEC) [7], RUSBoostClassifier (RBC) [12], UnderBaggingClassifier (UBC) (Maclin, 1997) 2.Methods using over sampling (generating new samples for minority classes): OverBoostClassifier (OBC), SMOTEBoostClassifier (SBC) [3], OverBaggingClassifier (OBAC) [11], SMOTEBaggingClassifier (SBAC) [16] 3.Reweighting-based ensembles (chaing the weight for different samples): AdaCostClassifier (ACC) [5], AdaUBoostClassifier (AUBC) [6], AsymBoostClassifier (ABC) [15].

2.3.2 Imbalanced Data Sets Classification Metrics

The most usual metrics precision and recall are accepted in this paper. Since the right recognition of the true values is more important for the analysis, so the value of recall is more meaningful.

Besides, based on some papers focusing on the imbalanced data set analysis, some metrics made for the imbalanced data analysis are also accepted: Cohen's Kappa (kappa), Youden's Index, Geometric Mean (g_mean), Matthew's Correlation Coefficient (MCC), Balanced Accuracy, F-Measure, lift, AFM, Area under the curve (AUC) [1].

In classification analysis, we usually evaluate a classifier by a confusion matrix (Table 3). In Table 3, the columns represent the classifier's predictions, and the rows are the actual classes. TP (True Positive) is the number of positive cases correctly classified as such. FN (False Negative) is the number of positive cases incorrectly classified as negatives. FP (False Positive) is the number of negative cases that are incorrectly identified as positive cases and TN (True Negative) is the number of negative cases correctly classified as such.

Table 2: Confusion matrix of two classes' classification.

	Classified positive	Classified negative
Actual positive	TP	FN
Actual negative	FP	TN

Table 3: Common performance measures based on confusion matrix.

Measure	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$

Cohen's Kappa(or Kappa)

$$kappa = \frac{accuracy - random\ accuracy}{1 - random\ accuracy} \quad (1)$$

where

$$random\ accuracy = \frac{(FP+TN)(TN+FN)+(FN+TP)(FP+TP)}{(TP+TN+FP+FN)^2} \quad (2)$$

The value of kappa ranges from -1 to +1. A value of 0 indicates there is no agreement between the actual and the prediction. A value of 1 means perfect concordance of the prediction result and the actual classes and a value of -1 indicates disagreement between classified classes and the actual ones.

Matthew's Correlation Coefficient

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3)$$

The Matthews correlation coefficient (MCC) is a correlation coefficient between the observed and predicted classifications. The value ranges from -1 to +1 where a value of +1 indicates a perfect prediction. 0 is no better than random prediction and -1 means the worst possible classification.

Adjusted F-Measure

$$AGF = \sqrt{F_2 \times InvF_{0.5}} \quad (4)$$

where

$$F_2 = 5 \times \frac{sensitivity \times precision}{(4 \times sensitivity) + precision} \quad (5)$$

After, the class labels of positive cases and negative cases are exchanged and

$$InvF_{0.5} = \frac{5}{4} \times \frac{sensitivity \times precision}{(0.5^2 \times sensitivity) + precision} \quad (6)$$

In this measure, the weight of patterns correctly classified the minority class is set a higher weight. A higher adjusted F-measure indicates a better-performing classifier. More metrics can be seen in Appendix.

The final performance of every algorithm is represented by all these indices, and the EasyEnsembleClassifier is selected as the best one for both part 1 and part 2 and the interpretation is performed on this algorithm.

2.4 Approaches to Feature Importance

2.4.1 Permutation Importance

The permutation importance of a feature is measured by the degree of reduction in the performance score of the model after randomly scrambling different feature values, which reflects the degree of dependence of the model performance on a particular feature [2]. If the final score of the model decreases significantly, the reliance on the feature is high; if the value of a feature has little effect on the model performance after randomization, the importance of the feature is low. In the experiments, by setting multiple parameters of random disruption, the scores of each feature in different random states can be derived and finally summarized.

2.4.2 SHAP

Not only to understand the importance of the features, but we also wonder how different values of one feature would affect the result. To this end, the SHapley Additive explanations (SHAP) approach is used [10]. The SHAP value is an arising machine learning model tool that can explain the detail of influence of a feature, based on the cooperative game theoretical concept Shapley value quantifying the contribution of each participant. And it can also be helpful to analyze the relationship of 2 variables in affecting the final result.

In machine learning experiments, we treat the explanatory variables as the players, and the v value function with respect to the given subset of variables is defined as follows [14]. Let f represent the machine learning algorithm that predicts the target variable (in this paper the final result of the art exam or of the enrollment) for a case x (in this paper a case is a d-dimensional vector containing the attributes of one record of a student). Let D be the index set of all features, i.e. $D = \{1, 2, \ldots, d\}$ and S be a subset of D. Then we set $f_S(x)$ as the conditional expectation of $f(x)$ using the values of the X_i features belonging to the set S, i.e.:

$$f_S(x) = \mathbb{E}(f(x)|X_i = x_i, \forall_i \in S) \quad (7)$$

Where if S is an empty set, then $f_S(x)$ is the expectation of $f(x)$, i.e. $f_{\{\}}(x) = E\big(f(x)\big)$. The expected values are calculated from the data. Then the value of a contribution of a subset of attributes is formally:

$$v(S) = f_S(x) - f_{\{\}}(x) \quad (8)$$

which is the change in result caused by observing the values of the S subset of features for a given case x. The contribution of the ith feature is calculated by its Shapley value corresponding to the $v_S(x)$ value function:

$$\varphi_i(x) = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!(d-|S|-1)!}{d!} \big(v(S \cup \{i\}) - v(S)\big) \quad (9)$$

3 RESULTS AND DISCUSSION

3.1 Prediction of The Result of The Art Exam

Firstly, the features whose permutation score is minus, which means that their influence on the result is unstable are excluded (Table 4). Considering the feature *Type of high school*, although it ranks the second in Figure 2, indicating it is the second significant feature, its permutation value of it is below zero. This means that whether the students who are better at main subjects are

more important than being in the art high schools can't be affirmed. Maybe more data about the students' academic performances in the 3 years in high school will help.

Secondly, through the modeling of the result of the art exam, we can see that the most important features are *Difficulty of art exam*, *Major*, and *Whether graduated from art high school* (Figure 2). For the feature *Difficulty of art exam*, we can see that the most difficult (6/7) majors have negative effects on the result, while the effects of other levels of difficulty are slightly positive (Figure3 a)). And the students coming from the art high schools will more likely apply for the most difficult majors such as Performing Arts. It is the SHAP methods telling us the details of the different effects of the different values of one feature. For the feature *Whether graduated from art high school*, this means that the training of skills and investment in advance is very critical, and the students who have determined to take the art related jobs before going to high school will have more advantages in the art exam, considering only 5% of the schools are art high schools. The features of Major (col_0, col_5, col_7) are mostly the new majors set by the institute of fine arts and the institute of Sounds such as new media design and special effects design. This is inspiring that students who just want to get easier to the BFA or haven't established a particular interest should better choose these new majors because the competitive pressure in these majors can be relatively small. Among the society-related features, *GDP per capita* is influential, which also relates to the investment in advance, and other features have little effect. From Figure3 b), we can see that when the value of GDP per capita is larger than 120000, it has positive effects on the art exam performance of one student, and meanwhile, it's more favorable for those students going to art high school.

Table 4: Average permutation importance of the most important features according to the SHAP value on predicting the result of the art exam.

Name	Permutation Importance
col_5	0.0130
col_0	0.0119
Difficulty of art exam	0.0046
Height	0.0038
GDP per capita	0.0034
Whether graduated from art high school	0.0031
col_7	0.0018
Type of sources of students (NGC)	0.0006
Type of sources of students (NGR)	-0.0045
Number of planned enrollments	-0.0058
Type of high School of Graduation	-0.0064
Number of colleges of project 985	-0.0085
Per capita disposable income	-0.0089

Table 5: Average performance of models in predicting the result of the art exam

	precision	recall	kappa	Youden_Index	g_mean	mcc	balanced_accuracy	f_measure	lift	AFM	AUC
SPEC	0.068	0.683	0.064	0.346	0.673	0.133	0.226	0.123	1.955	0.288	0.716
BCC	0.068	0.662	0.064	0.335	0.668	0.130	0.223	0.123	1.958	0.288	0.721
BRFC	0.069	0.683	0.066	0.350	0.675	0.135	0.228	0.125	1.980	0.286	0.723
EEC	0.076	**0.721**	0.079	**0.405**	**0.702**	0.158	**0.247**	0.137	2.184	0.304	**0.762**
RBC	0.086	0.560	0.095	0.347	0.664	0.152	0.220	0.150	2.486	0.322	0.721
UBC	0.081	0.592	0.087	0.350	0.670	0.147	0.224	0.142	2.332	0.313	0.737

OBC	0.088	0.587	**0.099**	0.368	0.677	**0.160**	0.229	**0.153**	2.538	0.326	0.728
SBC	0.086	0.416	0.091	0.258	0.592	0.126	0.175	0.143	2.488	0.320	0.691
OBAC	0.123	0.062	0.061	0.046	0.248	0.065	0.031	0.083	3.547	0.321	0.607
SBAC	0.140	0.073	0.073	0.056	0.267	0.078	0.036	0.095	**4.025**	**0.343**	0.618
AUBC	0.086	0.593	0.096	0.367	0.677	0.157	0.229	0.150	2.481	0.322	0.732
ABC	**0.152**	0.038	0.048	0.030	0.195	0.060	0.019	0.061	4.365	0.308	0.733

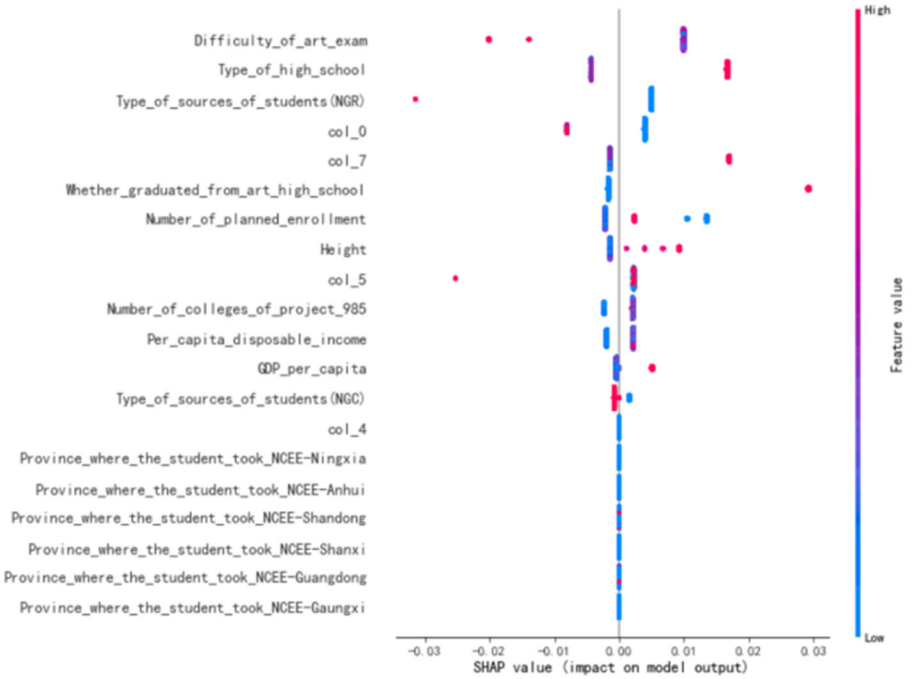
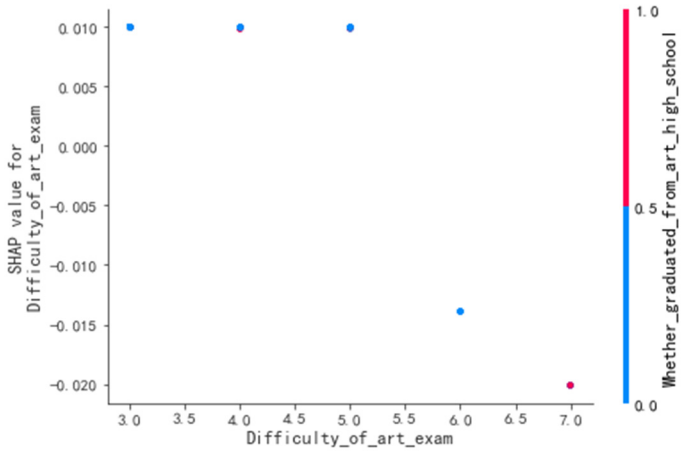
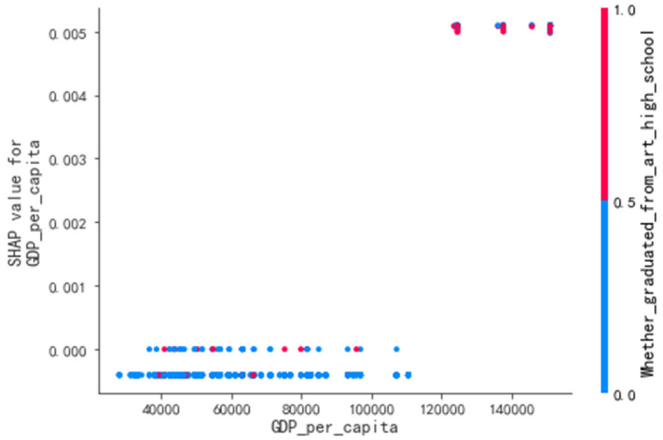


Figure 2: A summary plot showing the distribution of every feature impact in the EEC model on the result of the art exam. The color represents the feature value (red represents high and blue represents low; values greater than 0 represents a positive effect and below 0 mean negatives). Each record (one student can have multiple records) has one dot in each row. The *x* coordinate of the dot shows the impact of that feature on the prediction.



a)



b)

Figure 3: a) Dependence plot showing the relationship between *Difficulty of art exam* and *Whether graduated from art high school* (the most relevant feature). b) Dependence plot showing the relationship between *GDP per capita* and *Whether graduated from art high school* (the 3rd relevant feature).

3.2 Prediction of Final Admission

Firstly, the features whose permutation score is minus, which means that they have little influence on the result are excluded as the same (Table 6). Then through the modeling and prediction of the result of the art exam, we can see that the most 3 important features in SHAP values and permutation values are the same: *Rankings in the major*, *Number of planned enrollment,* and *Score of art exam* (Figure 4). This means that for art schools, the result of the art exam is very important in deciding whether a student will be admitted.

Then among the features that related to the score of the NCEE, *Number of colleges*, *Number of planned enrollments of all colleges* and *Age* is also very significant. It proves that in provinces with a higher number of colleges and a higher number of planned enrollments, the data related to students' NCEE scores or ranking will be a bit more favorable because there is less competition there. And more students in these provinces will choose their local schools (Figure 5 a)). There may be some deeper reasons to explore more about the emergence of a few specific provinces like Shaanxi and Jiangxi etc., but the rankings of these characteristics are extremely low, and their permutation values are very close to 0. In fact, it can be boldly assumed that these characteristics are almost negligible. Same with the feature *col_3*. It also indicates that in the BFA, although there are different NCEE score requirements among the majors and the cut-off points of many provinces are different, the influence of different provinces remains nearly the same.

Secondly, in the prediction of the results of the two exams, we can see that among the society-related features, the *GDP per capita* has more influence on the art exam, and the information related to the colleges and universities in each province (representing a certain level of education) is more important to the NCEE and the final result. Namely that the pre-training of the skills related to the art exam and the preparation of related works needs more financial investment, and the influence of financial investment on the NCEE score and the final result is negligible. This is consistent with the research [13].

Table 6: Average permutation importance of the most important features according to the SHAP value on predicting the result of the final enrollment.

Name	Permutation Importance
Rankings in the major	0.1474
Number of planned enrollments	0.0493
Score of art exam	0.0061
Age	0.0058
Number of planned enrollments of all colleges	0.0046
col_3	0.0025

Number of colleges	0.0020
Student's preference order for the major	0.0020
Province where the student took NCEE - Shaanxi	0.0018
Household registration location - Hubei	0.0015
Household registration location - Jiangxi	0.0005
Type of sources of students (OGR)	-0.0010
Rankings in the major within the same province	-0.0015
col_6	-0.0020

Table 7: Average performance of models in predicting the result of the final enrollment.

	precision	recall	kappa	Youden_ Index	g_mean	mcc	balanced_accuracy	f_measure	lift	AFM	AUC
SPEC	0.512	0.683	0.299	0.322	0.660	0.309	0.218	0.585	1.434	0.576	0.716
BCC	0.515	0.683	0.303	0.326	0.662	0.312	0.219	0.587	1.442	0.578	0.721
BRFC	0.503	0.715	0.294	0.322	0.659	0.309	0.217	0.590	1.407	0.574	0.702
EEC	0.558	0.619	**0.338**	**0.346**	**0.671**	**0.339**	**0.225**	0.587	1.562	0.592	**0.737**
RBC	**0.580**	0.530	0.323	0.317	0.646	0.324	0.209	0.554	1.624	0.569	0.699
UBC	0.537	0.648	0.323	0.337	0.668	0.326	0.223	0.587	1.504	0.584	0.718
OBC	0.568	0.505	0.299	0.292	0.630	0.300	0.199	0.535	**1.591**	0.553	0.696
SBC	0.473	**0.804**	0.265	0.306	0.635	0.300	0.202	**0.596**	1.324	**0.598**	0.719
OBAC	0.543	0.520	0.279	0.276	0.627	0.279	0.197	0.531	1.520	0.548	0.704
SBAC	0.560	0.548	0.310	0.309	0.646	0.310	0.208	0.554	1.568	0.549	0.711
ACC	0.572	0.552	0.325	0.322	0.652	0.325	0.213	0.562	1.602	0.563	0.690
AUBC	0.572	0.552	0.325	0.322	0.652	0.325	0.213	0.562	1.602	0.563	0.699
ABC	0.572	0.552	0.325	0.322	0.652	0.325	0.213	0.562	1.602	0.563	0.699

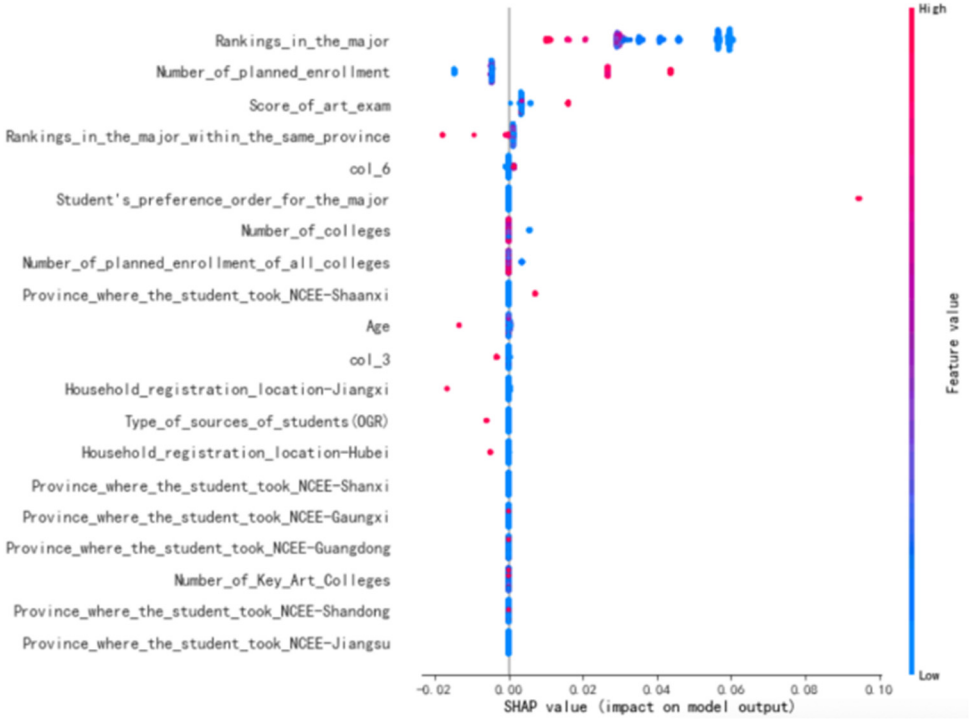
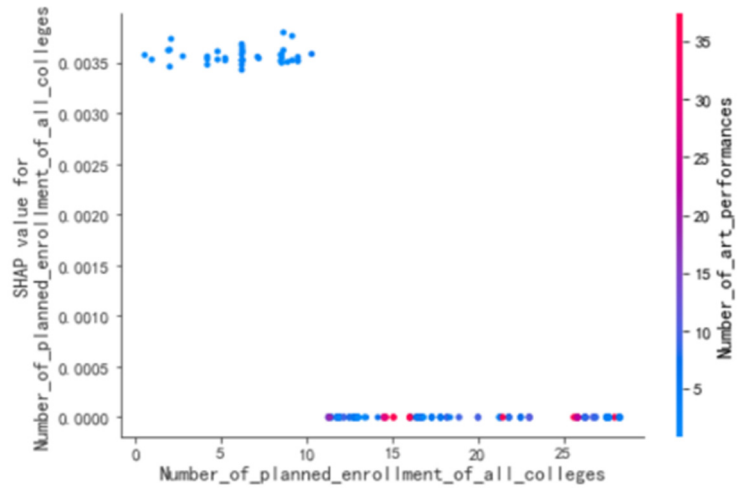
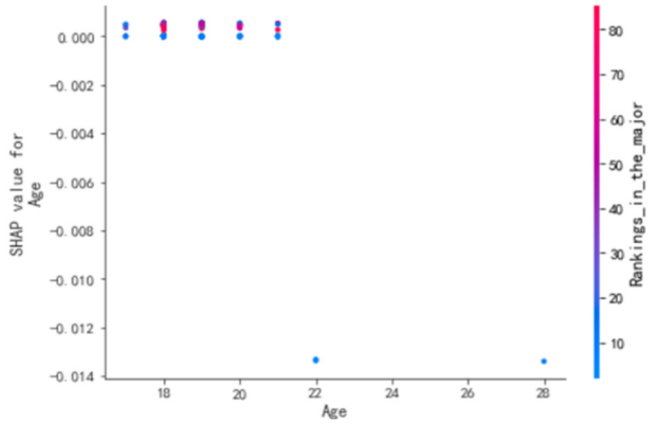


Figure 4: A summary plot showing the distribution of every feature impact in the EEC model on the result of the final enrollment. The color represents the feature value (red represents high and blue represents low; values greater than 0 represents a positive effect and below 0 mean negatives). Each record (one student can have multiple records) has one dot in each row. The x coordinate of the dot shows the impact of that feature on the prediction.



a)



b)

Figure 5: a) Dependence plot showing the relationship between *Number of planned enrollments of all colleges* and *Number of art performances* (the most relevant feature). b) Dependence plot showing the relationship between *Age* and *Rankings in the major*. Students who are older than 22 will be in a slightly unfavorable situation.

4 CONCLUSIONS

In this work, we follow the flow of data science, from collecting data, doing the feature engineering, applying the proper intelligent algorithms, and finally evaluating and explaining the results. The excellent performances of the new-added features, IMBENS, and SHAP in the interpretation of the result are noticed.

By analyzing the 3-year data of the BFA, many conclusions about the art exam and admission to the art schools can be seen. For the students willing to apply for the art schools, the plentiful pre-training of the skills and works and the proper selection of majors are very important. For art schools, the major settings and the difficulty of the exam is very critical, teachers must take more consideration when writing the questions and setting the standards. And the promotion of the new majors needs some work, otherwise, students won't know the skills needed by them and the future of the majors.

The data used in this work is limited, and many students' related data such as the performance of the student in the high school, and the financial background of the family are hard to get. We wish that this work can be inspiring for those researchers who aim for analyzing the art exam or the admission standards.

APPENDIX

Youden's Index

$$Youden's\ index(\gamma) = sensitivity - (1 - specificity) \quad (10)$$

Geometric Mean(G-Mean)

$$G - mean = \sqrt{sensitivity \times specificity} \quad (11)$$

Balanced Accuracy

$$Balanced\ Accuracy = \frac{1}{2}(sensitivity \times specificity) \quad (12)$$

F-measure

$$F - Measure = \frac{2 \times sensitivity \times precision}{sensitivity + precision} \quad (13)$$

ACKNOWLEDGEMENTS

It should be acknowledged that this work was supported by the research project of undergraduate education and teaching of Beijing Film Academy (grant number XYJS202013).

REFERENCES

[1] Akosa, J. (2017, April). Predictive accuracy: A misleading performance measure for highly imbalanced data. In Proceedings of the SAS Global Forum (Vol. 12).

[2] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340-1347.

[3] Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003, September). SMOTEBoost: Improving prediction of the minority class in boosting. In European conference on principles of data mining and knowledge discovery (pp. 107-119). Springer, Berlin, Heidelberg.

[4] Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley, 110(1-12), 24.

[5] Fan, W., Stolfo, S. J., Zhang, J., & Chan, P. K. (1999, June). AdaCost: misclassification cost-sensitive boosting. In Icml (Vol. 99, pp. 97-105).

[6] Karakoulas, G., & Shawe-Taylor, J. (1998). Optimizing classifers for imbalanced training sets. Advances in neural information processing systems, 11.

[7] Liu, X. Y., Wu, J., & Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2), 539-550.

[8] Liu, Z., Cao, W., Gao, Z., Bian, J., Chen, H., Chang, Y., & Liu, T. Y. (2020, April). Self-paced ensemble for highly imbalanced massive data classification. In 2020 IEEE 36th international conference on data engineering (ICDE) (pp. 841-852). IEEE.

[9] Liu, Z., Wei, Z., Yu, E., Huang, Q., Guo, K., Yu, B., ... & Chang, Y. (2021). IMBENS: Ensemble Class-imbalanced Learning in Python. arXiv preprint arXiv:2111.12776.

[10] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[11] Maclin, R., & Opitz, D. (1997). An empirical evaluation of bagging and boosting. AAAI/IAAI, 1997, 546-551.

[12] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 40(1), 185-197.

[13] Shen. (2018). A Study about the Influence of Family Economic Capital on the Educational Status Attainment of Art Examinees: Based on a Survey of F University (Master's thesis, Fuzhou University). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202001&filename=1019096588.nh

[14] Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems, 41(3), 647-665.

[15] Viola, P., & Jones, M. (2001). Fast and robust classification using asymmetric adaboost and a detector cascade. Advances in neural information processing systems, 14.

[16] Wang, S., & Yao, X. (2009, March). Diversity analysis on imbalanced data sets by using ensemble models. In 2009 IEEE symposium on computational intelligence and data mining (pp. 324-331). IEEE.

[17] Zhang. (2017). A Study on the Balanced Development of Art Students' Culture and Professional Background(Master's thesis, Hunan Normal University).https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202001&filename=1017133063.nh