PART ONE

Predictive variable processing

Multi-arts program viewing forecast variable processing - dependent variable

- **Sample: 139 completed quarterly programmes in the Ari database from January 2017 to October 2018.**

- Broadcast volume: Average unicast volume during each broadcast period, ensuring that programs with different broadcast cycles are still comparable.

- Unicast volume is divided into five levels according to platform rules and average traffic distribution:

Programme classification	Number of programmes
S + (32M +)	6
S (12M +)	17
Grade A (8M - 12M)	15
Class B (4M-8M)	31
Class C (4M-)	70

Comprehensive Arts Programs View Forecasting Variable Processing - An Overview of Independent Variables

- Sample: 139 completed quarterly programmes in the Ari database from January 2017 to October 2018.

Variable	Variable interpretation
Type of programme	139 programs were classified into 10 categories and the programme types were classified as variables.
Video/Net	Divide the program into TV and net by playing platform, which is a classification variable
Solo/jigsaw	Program broadcast on only on one video platform, or on multiple video platforms, this variable is classification variable.
Playback platform	Because the number of users of the webcast video platform is different, the influence is also different. Variables defined the platform with most play counts.
Production cost	Programme production costs, data provided by Simei
Promotional costs	Programme publicity costs, data provided by Simei
Star influence	Star influence variable is to calculate and sum up the Baidu index for the fixed guests and the first guests 3-6 months before the program is broadcast.
Playback period	According to the broadcast time of different programs, the broadcast time is divided into three stages: Summer, year and hour. The variable is a classification variable.
Integrated N generation	Does the show have a sequel (0-1 variable)
Competition factor	Number of programmes of the same type that are broadcast weekly during a programme

Source: Ari consulting database

Analysis of Reception Forecasting Variable Processing - Continuous Variable Correlation

- Production costs, publicity costs, star influence and broadcasting volume have positive correlation, competition factors and broadcasting volume have negative correlation.

Continuous Variable and Dependent Variable Correlation Matrix

	Production cost	Promotional costs	Star influence	Competition factor	Average amount of play
Production cost	1.00	0.58	0.19	-0.11	0.43
Promotional costs	0.58	1.00	0.31	-0.21	0.50
Star influence	0.19	0.31	1.00	-0.12	0.46
Competition factor	-0.11	-0.21	-0.12	1.00	-0.11
Average amount of play	0.43	0.50	0.46	-0.11	1.00

Source: Ari consulting

Analysis of Variable Processing in Comprehensive Arts Program's Reception Forecasting

- Categorized variables cannot see their relevance to dependent variables, so the influence of a particular variable can be seen through the average amount of play corresponding to different categorized variables.

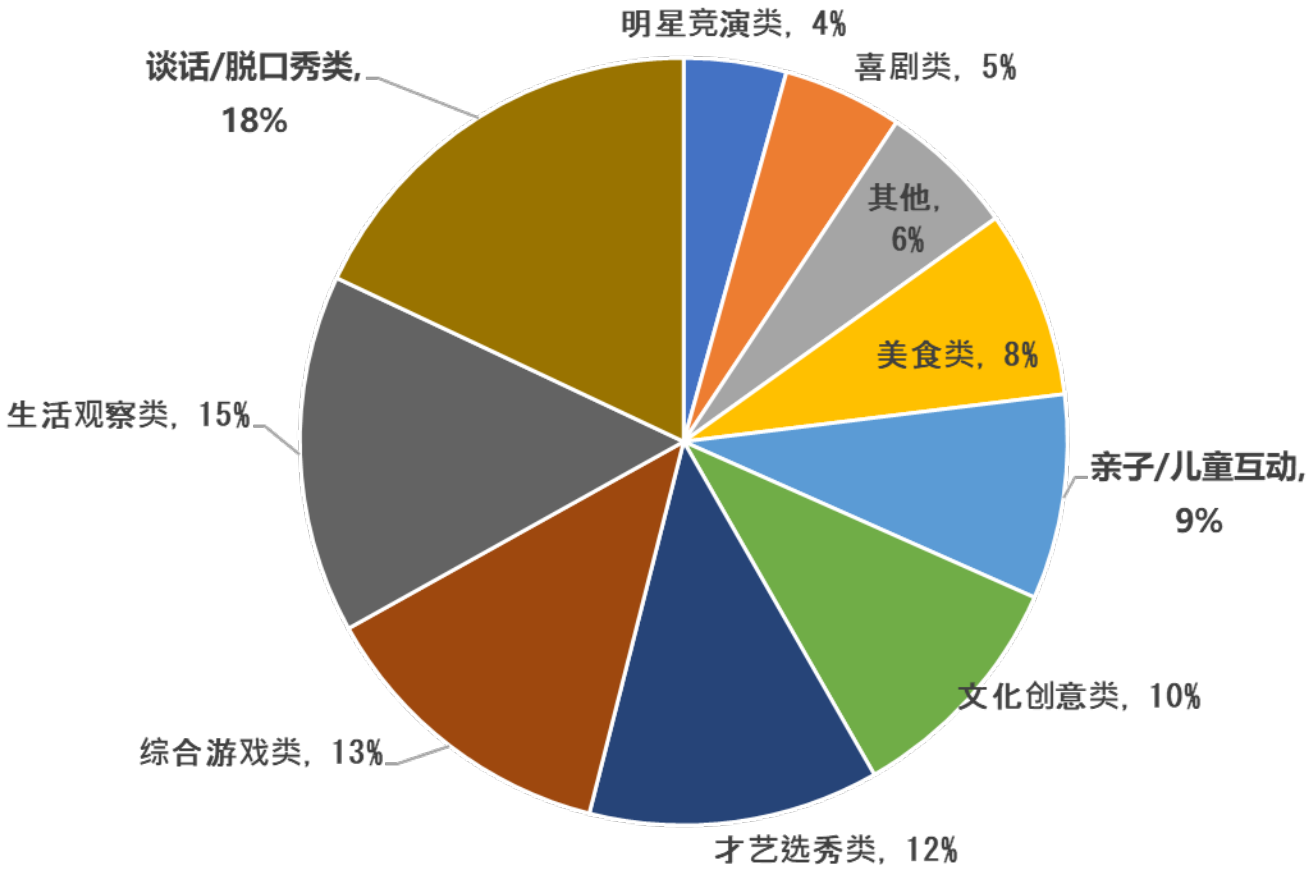
Variable	Specific variable	Average amount of play
Video network	Television versatility	8,225, 220
	Net versatility	7,081, 126
Integrated N generation	Non-integrated N generation	5,412, 720
	Integrated N generation	10,603, 810
Solo/jigsaw	Solo	7,912, 090
	Jigsaw broadcasting	6,932, 616
Playback period	Normal	7,092, 583
	Summer stalls	10,014, 690
	Year of celebration	5,131, 215
Playback platform	Idyllic art	11,644, 270
	Tencent video	5,993, 272
	Mango TV	5,320, 887
	Yorkie	4,297, 073
	Sohu video	8 75, 894

Variable	Specific variable	Average amount of play
Type of programme	Talent selection	20,120, 320
	Integrated games	12,908, 280
	Star rehearsal class	7,836, 497
	Comedy	6,925, 319
	Gourmet	6,113, 874
	Life observation class	5,250, 686
	Parent-child/child interaction	4,706, 169
	Other	4,612, 079
	Talk/talk show	3,851, 858
	Cultural creativity	1,540, 569

Source: Ari consulting database

Multi-arts programme viewing forecast variable processing - programme type

- 139 programs were classified into 10 categories and the programme types were classified as variables.



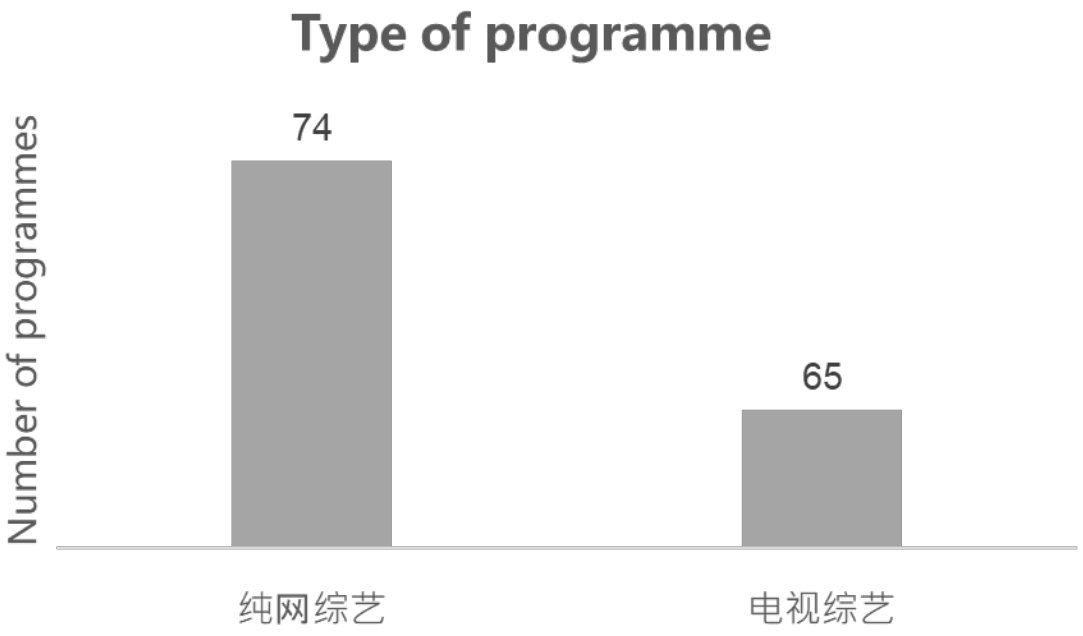
Note: Other programs include scientific knowledge and machine sports.

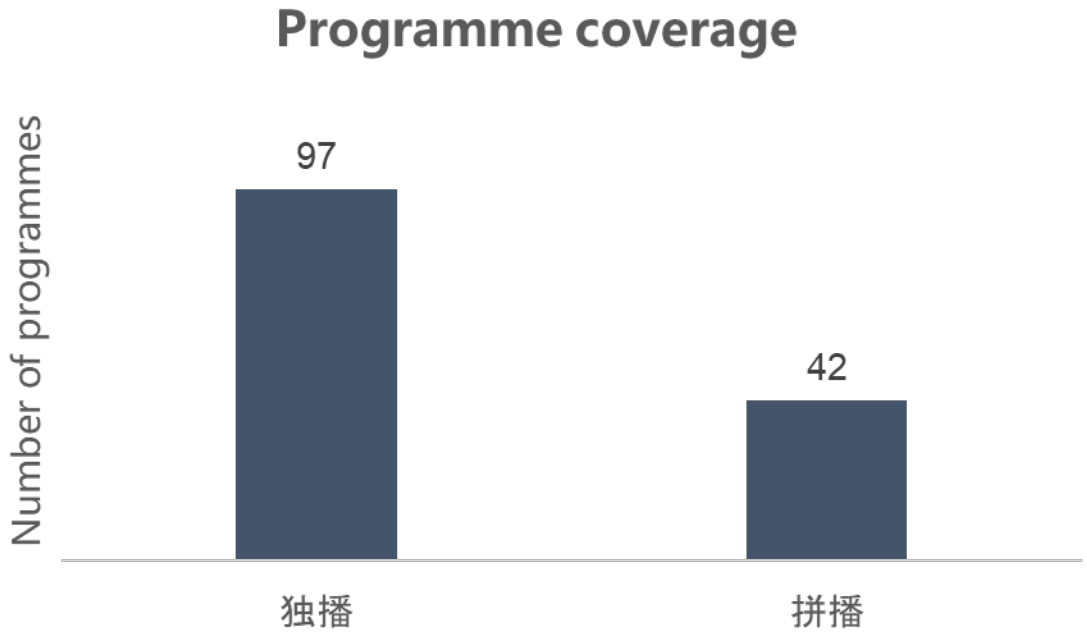
Type of programme	Classification rule
Life observation class	No theme is set, the rhythm is relaxed, the guest's life is observed through the camera, thus getting the sense and understanding of life.
Talent selection	There are a variety of standard selection programs, the players are chiefly vegetarians.
Talk talk show	Guests (stars/vegetarians) gathered to discuss a topic/theme
Cultural creativity	It involves culture, history, science and technology and so on.
Integrated games	Each game with different themes and tasks requires guest competition, including indoor and outdoor games.
Gourmet	Food production/sharing/conversations/perceptions
Parent-child/child interaction	Programmes around children, including adults/children; Children/children, children/pets interaction, etc.
Star rehearsal class	Unlike talent shows, the players are stars
Comedy	Comedy/cross talk programmes
Other	Other, including fashions, etc.

Source: Ari consulting database

TV Forecast Variable Processing
- Video/Net & Broadcast/Multicast

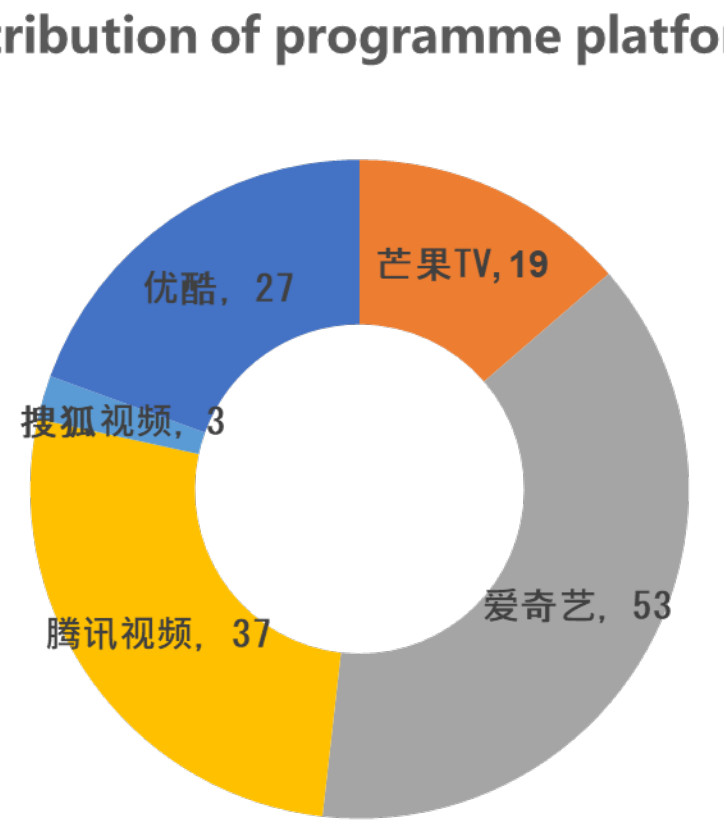
- Video/Networking: Divide the programme into video and net by broadcasting platform, which is classified as variable
- Solo/jigsaw: At present, versatile arts are basically broadcast on a web-based platform. Programs broadcast on only one video platform are solo programs. Programs broadcast on multiple video platforms are jigsaw programs.

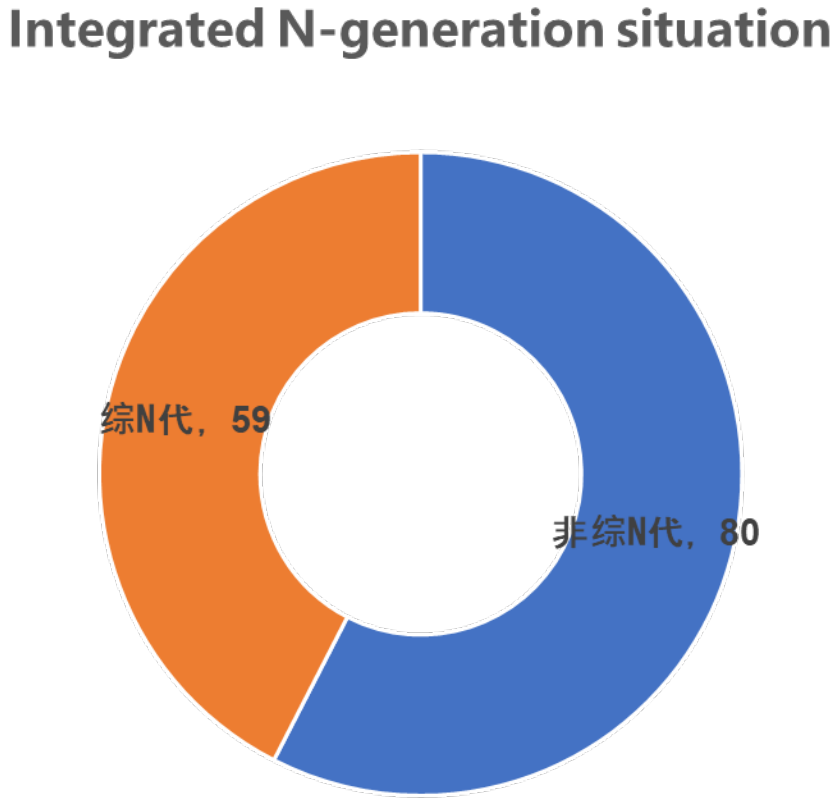




Comprehensive Arts Programs View Forecast Variable Processing - Play Platform & Integrated N Generation

- Playing Platform: The broadcasting platform takes the platform with the highest broadcasting volume as a variable.
- Integrated N Generation: Sequel to Play Play





Multi-arts programme viewing forecast variable processing - production costs & publicity costs

- Production costs: Programme production costs
- Promotional costs: Programme promotional costs

Examples

Programme name	Production costs (10,000 yuan)	Publicity costs (10,000 yuan)
Run	15, 000	800
New Chinese rap	150, 000	1, 000
Good Chinese Voice	8, 000	1, 000
The Second Quarter of Chinese New Singing	9, 000	800
Run the second season	150, 000	1, 000
China has hip-hop	25, 000	1, 000
Creation 101	110, 000	1, 000
Warm Blood Street Dance	140, 000	1, 000
Idol practitioner	9, 000	800
Second quarter of the Tank Conference	3, 500	50

Multifaceted Programs Reception Forecasting Variable Processing - Playback Period & Competition Factors

- Programme schedule: The schedule is divided into three stages according to the schedule, as shown in the table below, and the schedule is classification variable.
- Competition factor: The number of programmes of the same type that are broadcast weekly during a programme

Programme schedule

Programme period	Programme schedule	Number of programmes broadcast during the period
Early July to early September	Summer stalls	39
End of November to early March of the following year	Year of celebration	21
Early March to late June, early September to late November	Normal	79

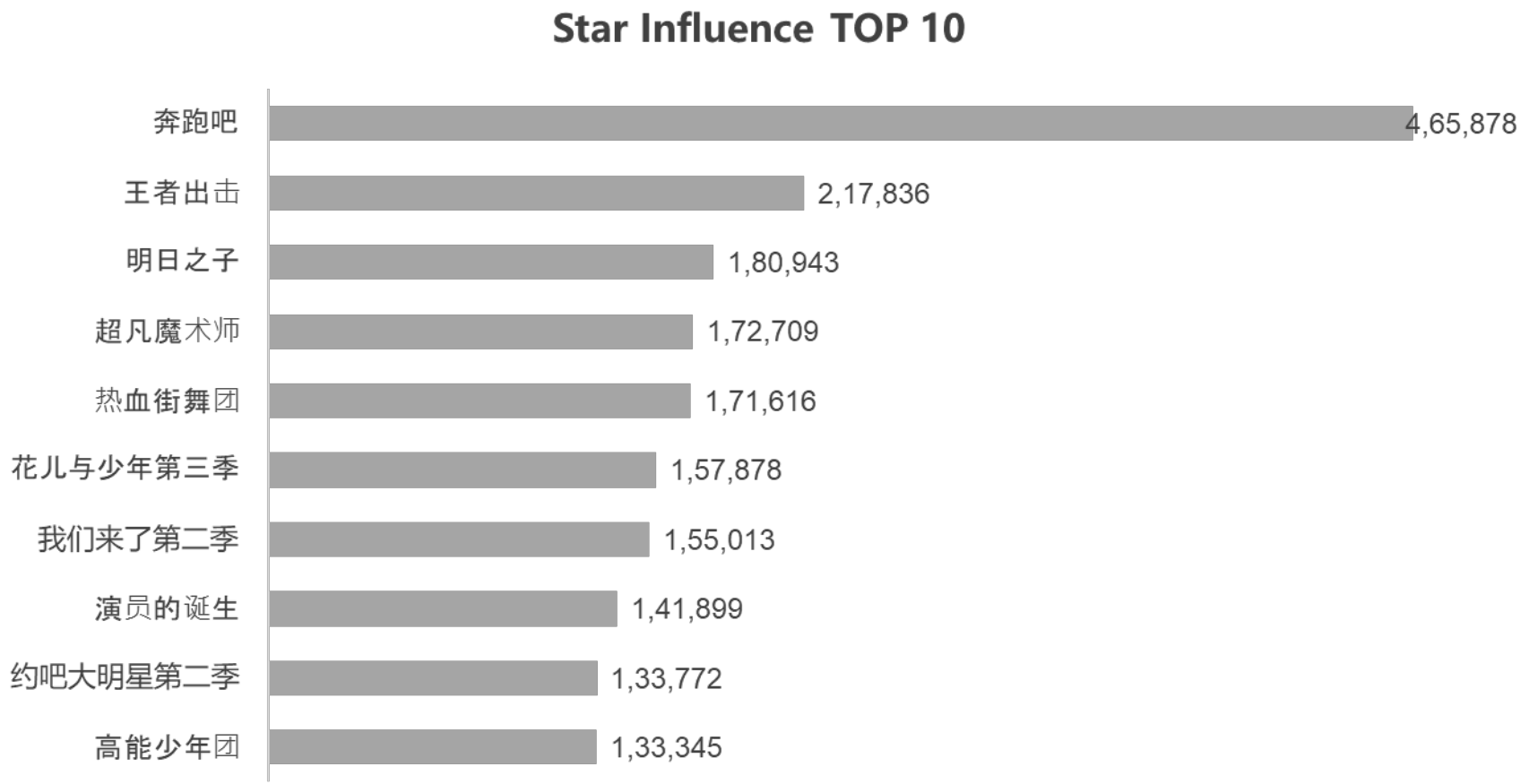
Number of similar performances per week during programme broadcast

Type of programme	Average number of programmes per week
Talent selection	3.4
Gourmet	1.9
Star rehearsal class	1.6
Other	2.6
Parent-child/child interaction	2.4
Life observation class	3.7
Talk/talk show	6.5
Cultural creativity	3.3
Comedy	1.4
Integrated games	3.3

Source: Ari consulting

Variety of Arts Programs Reception Forecast Variable Handling - Star Influence

- Star Guest Factor: Since it is generally only known when engaging in versatile arts, the Star Guest Factor Variable is calculated and summed up by calculating the Baidu Index of the fixed guest and the first guest 3-6 months before the programme is broadcast.



PART TWO

Program Reception Forecasting Model

The Reception Forecast of Comprehensive Arts Programs-Stochastic Forest Model

Modeling process

- Classification variable conversion
- Continuous variable processing

- SMOTE algorithm
- Data set cutting

- Tuning
- Evaluation

The Reception Forecast of Comprehensive Arts Programs-Stochastic Forest Model

Data Processing & Data Optimization

- **Data processing**
  - **Classification variable conversion**
    Video network integration (0-1 variable), programme type (classification variable), playback platform (classification variable), playback time (classification variable), integrated N generation (0-1 variable), unicast (0-1 variable)
  - **Continuous variable processing:**
    Discretization of continuous variables, which are cut into four categories by the quartile of variables - Production costs, publicity costs, star influence, competition factors

- **Data optimization**
  - **SMOTE algorithm:** The number of programs is unbalanced in different levels. From the training model point of view, if the number of samples in a class is small, the "information" provided by this class is too small. Therefore, SMOTE over-sampling algorithm is used to deal with the data imbalance, and a new sample is synthesized based on "interpolation".

Programme classification by broadcast volume

Programme classification	S+	S	A	B	C
Number of programmes	6	17	15	31	70

Program Distribution Processed by SMOTE Algorithm

Programme classification	S+	S	A	B	C
Number of programmes	70	70	70	70	70

- **Data set cutting:** Cut data into test and training sets in a 1: 4 ratio

The Reception Forecast of Comprehensive Arts Programs-Stochastic Forest Model

Model optimization

Random forest model

In machine learning, a random forest is a classifier that contains multiple decision trees, and the class output depends on the number of classes output by individual trees.

- **Optimization of random forest models:**

  GridSearchCV is used to adjust the parameters automatically. If the parameters are entered, the optimal results and parameters can be given. Determine the maximum number of weak learners (that is, the number of decision trees n estimators) and the maximum depth of a single class decision tree (max _ depth)

N_estimators	Max_depth
80	8

Variable importance, 10 variables sorted by importance:

From the importance score, it can be seen that the competition factor is the most important factor that affects the play quantity classification, and whether the program is broadcast alone or not is net-only comprehensive to have little influence on the play quantity.

Independent variable	Importance score
Competition factor	0.15
Production cost	0.13
Playback platform	0.13
Promotional costs	0.12
Type of programme	0.12
Star influence	0.09
Playback period	0.09
Integrated N generation	0.07
Solo or not	0.05
Video network	0.04

The Reception Forecast of Comprehensive Arts Programs-Stochastic Forest Model

Model evaluation

- **Model assessment:**

  The 5-fold cross-validation of the data was carried out, that is, the data were averagely divided into 5 parts, one of which was used as the test data and the other as the training data to calculate the accuracy of the five models: (0.79, 0.86, 0.89, 0.88, 0.96) All the accuracy was greater than 0.79

- **Confusion matrix:**

  Split the test set and training set according to the ratio of 1: 4 to obtain the confusion matrix of the test set. Confusion matrix is also called error matrix, which is a standard form of accuracy evaluation. In the Confusion matrix of the following test set, the first column represents the prediction for category A. It was found that 11 of 14 A-type samples were correctly classified, 2 were misclassified into C and 1 was misclassified into S (by column). So the diagonal of the confusion matrix represents the correct number of classifications.

		Forecast programme classification				
		A	B	C	S	S+
Real programme classification	A	**11**	1	2	0	0
	B	0	**14**	1	0	1
	C	2	4	**12**	0	0
	S	1	1	0	**10**	0
	S+	0	0	0	0	**9**

Evaluation indicators for classification results:

Precision = (Number of samples predicted to be 1 and correctly predicted)/(Number of samples predicted to be 1 but incorrectly predicted.) = TP/(TP + FP)

Recall = (number of samples predicted to be 1 and correctly predicted)/(number of samples to be predicted 0 but incorrectly predicted) = TP/(TP + FN)

TP: Forecast is 1 (Positive), actual is also 1 (Truth-forecast is right)

TN: 0 (Negative), actually 0 (Truth-right)

FP: Forecast 1 (Positive), Actual 0 (False - Forecast wrong)

FN: 0 (Negative), 1 (False - false)

The following table can be obtained from the test set obfuscation matrix:

	Precision	Recall	F1-score	Support
S+	0.79	0.79	0.79	14
S	0.70	0.88	0.78	16
A	0.80	0.67	0.73	18
B	1.00	0.77	0.87	13
C	0.82	1.00	0.90	9
MICRO	0.80	0.80	0.80	70
MACRO	0.82	0.82	0.81	70
WEIGHTED	0.81	0.80	0.80	70

Source: Ari consulting database

Appendix

• SMOTE algorithm

SMOTE (Synthetic Minoritye over-sampling technique, SMOTE) is an oversampling algorithm proposed by Chawla in 2002, which can avoid the above problems to some extent.

- Many classification problems usually face the problem of sample disequilibrium, and many algorithms are not ideal in this case. To solve the unbalanced problem, there are two strategies: Sampling and cost-sensitive learning. Smote algorithm is a common one in over-sampling.
- The idea of smote algorithm is to synthesize new minority samples. The strategy of synthesis is to randomly select a sample b from its nearest neighbor for each minority sample a, and then randomly select a point on the line between a and b as a newly synthesized minority sample.

• Cross-validation and grid search:

Cross validation is a generalization of data sets independent of training data, which can avoid overfitting problems.

(1) The number of training sets should be large enough, generally more than half

2) Uniform sampling of training set and test set

Grid Search: A method of adjusting parameters for exhaustive search. In all candidate parameter choices, the best parameter to perform is the final result of trying every possibility through cyclic traversal. The principle is like finding the maximum in an array.

Why is it called a grid search?

Taking a model with two parameters as an example, there are three possibilities for parameter a and four possibilities for parameter b to list all possibilities, which can be represented as a 3 * 4 table, where each cell is a grid and the cycle is like traversing and searching in each grid, so called grid search