Intelligent Modeling Introduction

AI Model

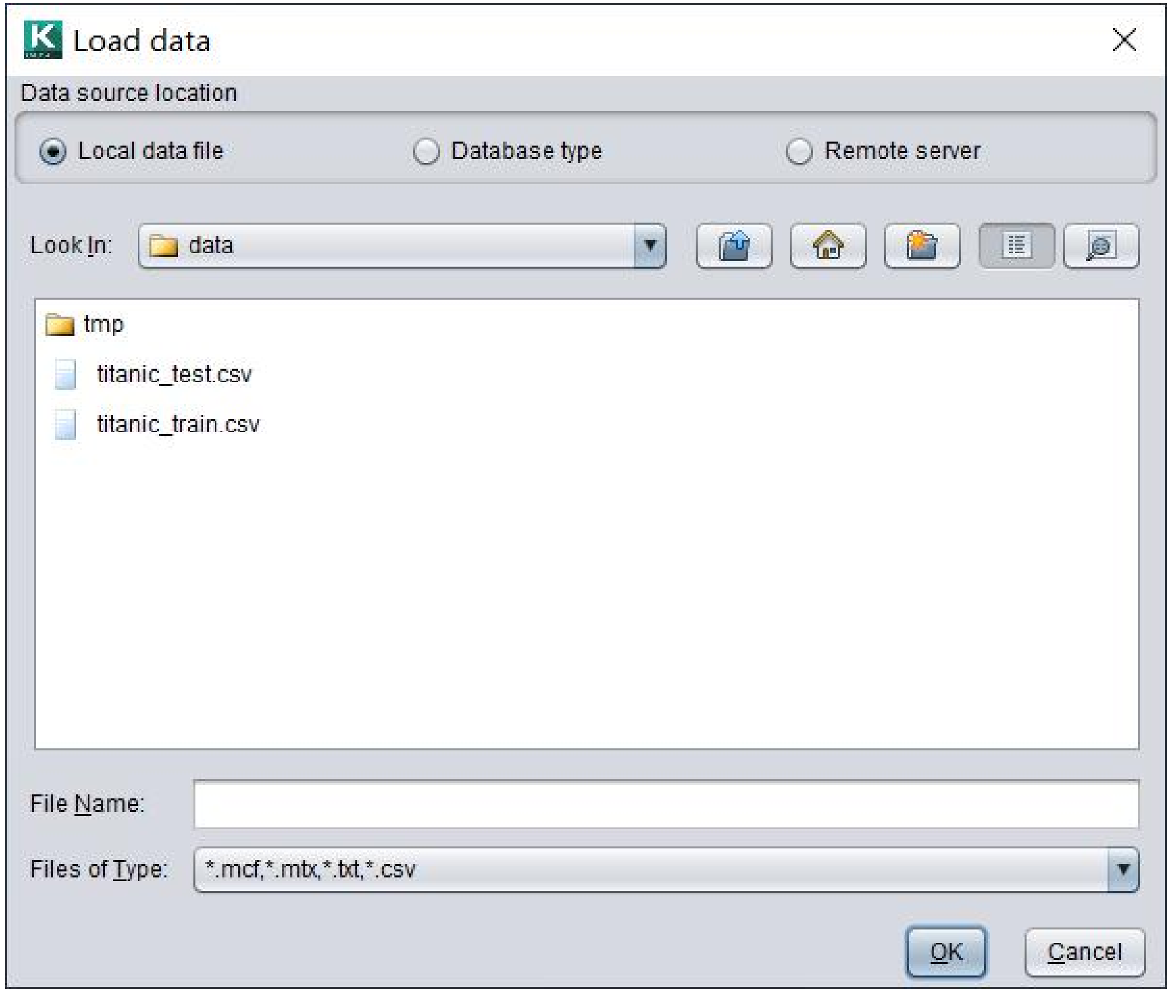
CONTENTS

CONTENTS

01 Data Source

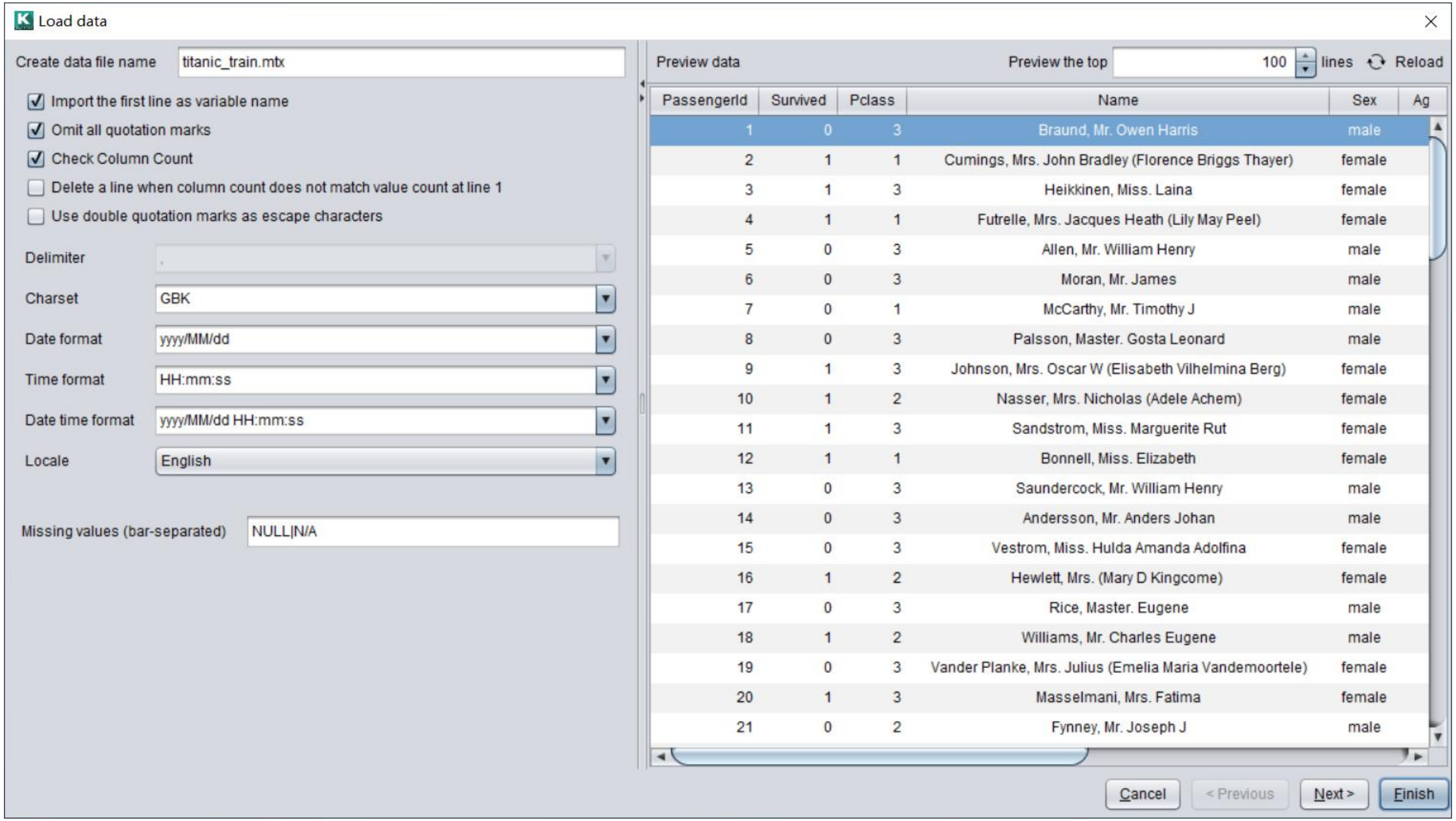
1. Local data file

Intelligent modeling supports TXT, CSV and other data files.



1. Local data file

After selecting a file, you can define the parameter configuration of the data file.

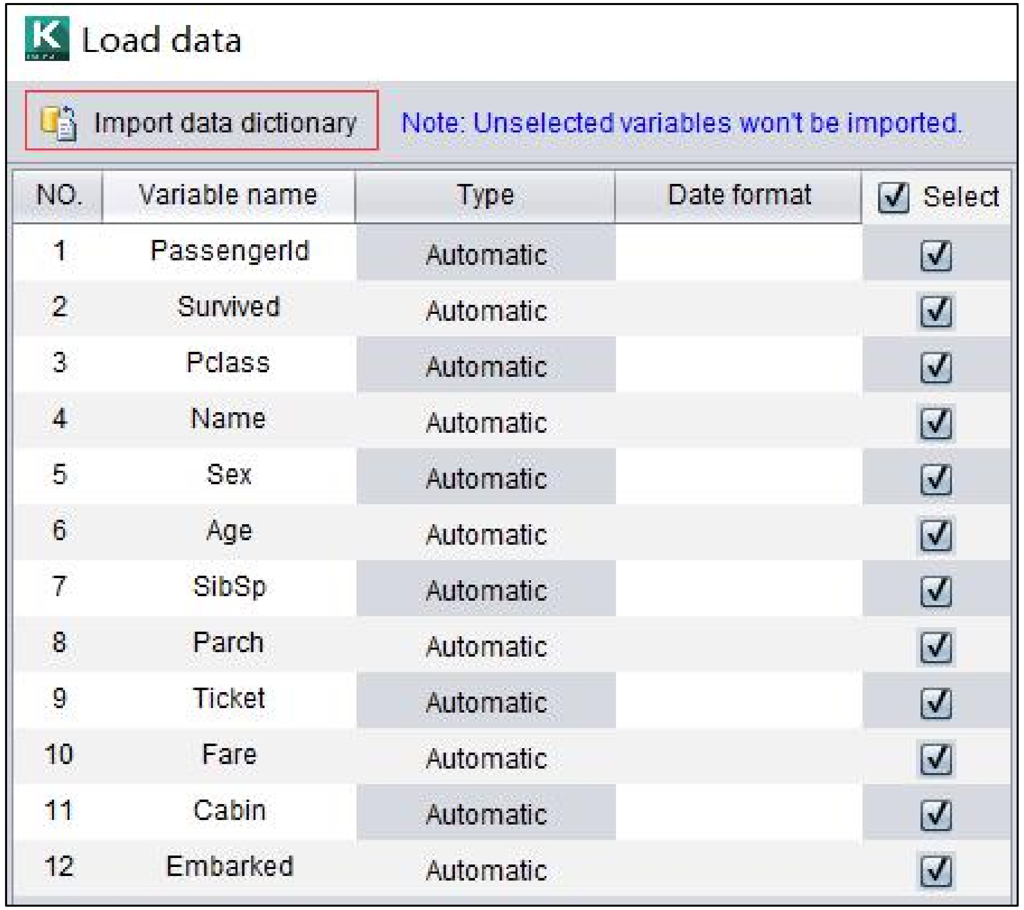


1. Local data file

Next, you can define the variable type, date format, and selection status.

Variable types can be automatically detected or be configured by importing the data dictionary.
The format of data dictionary is as follows:

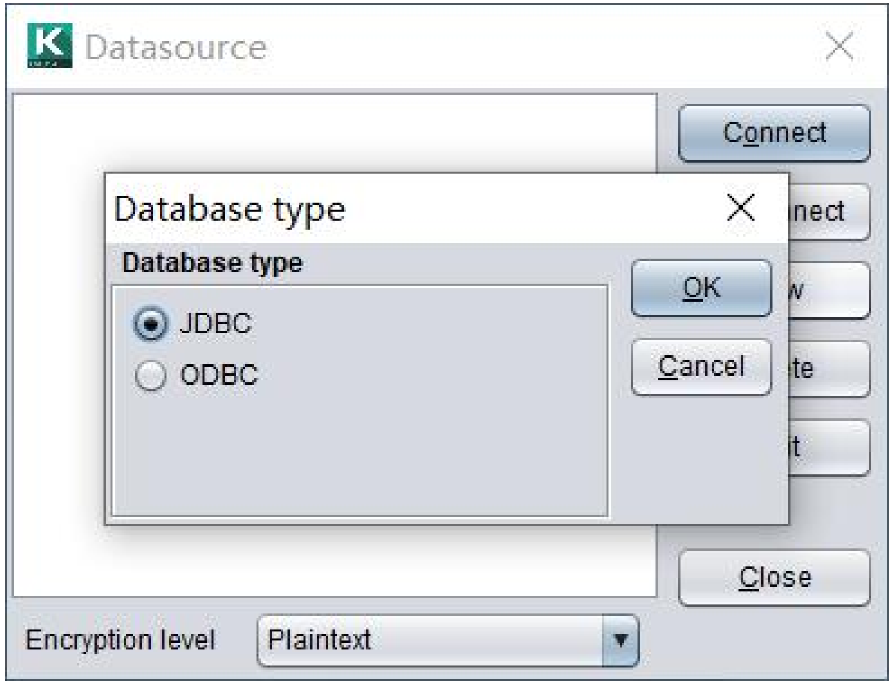
Name	Type	DateFormat	Used	Importance
PassengerId	Identity		TRUE	0
Survived	Binary		TRUE	0
Pclass	Categorical		TRUE	0
Name	Text		FALSE	0
Sex	Binary		TRUE	0
Age	Numerical		TRUE	0
SibSp	Categorical		TRUE	0
…	…	…	…	…

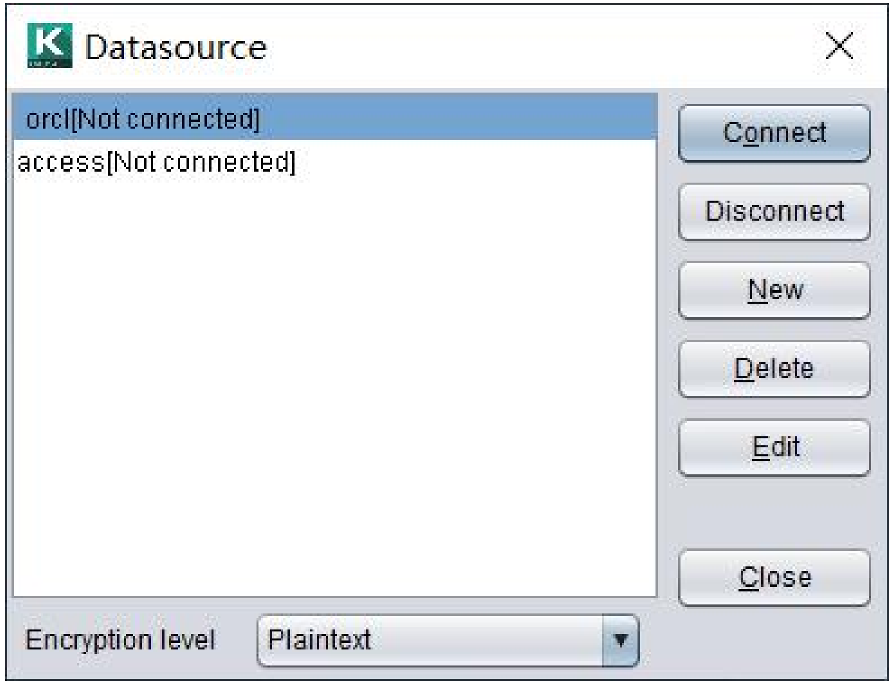


NO.	Variable name	Type	Date format	Select
1	PassengerId	Automatic		☑
2	Survived	Automatic		☑
3	Pclass	Automatic		☑
4	Name	Automatic		☑
5	Sex	Automatic		☑
6	Age	Automatic		☑
7	SibSp	Automatic		☑
8	Parch	Automatic		☑
9	Ticket	Automatic		☑
10	Fare	Automatic		☑
11	Cabin	Automatic		☑
12	Embarked	Automatic		☑

2. Database

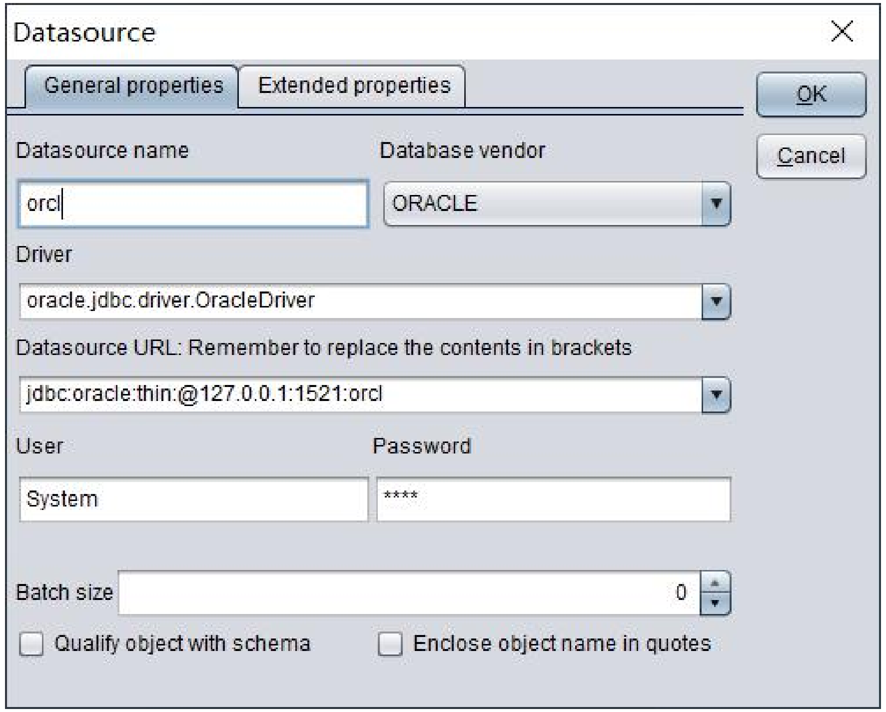
In the data source window, you can define two data source connections: JDBC and ODBC.



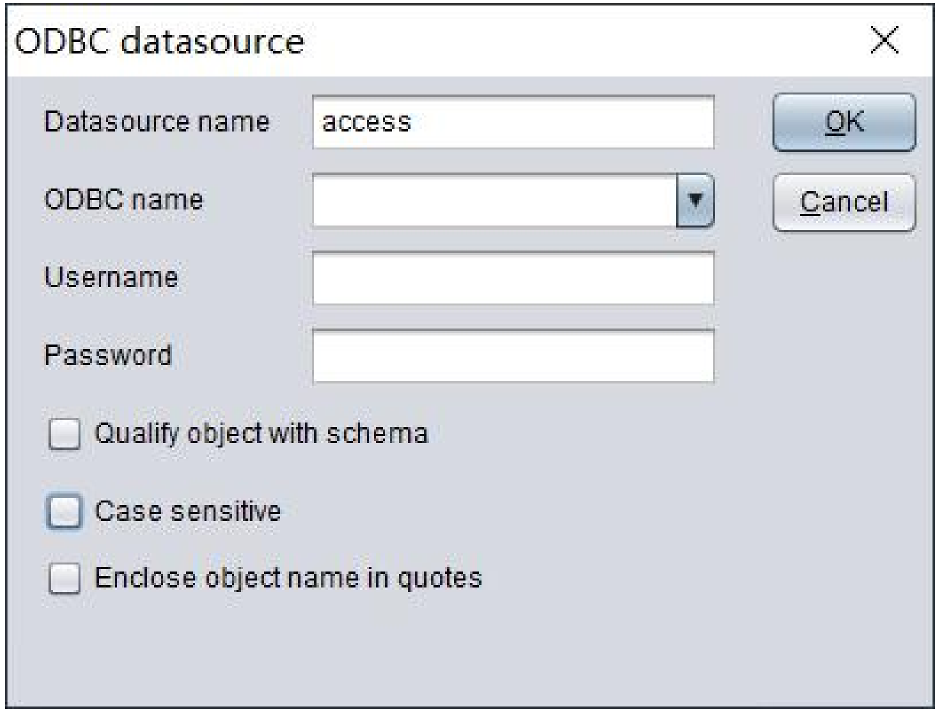


2. Database

JDBC Datasource

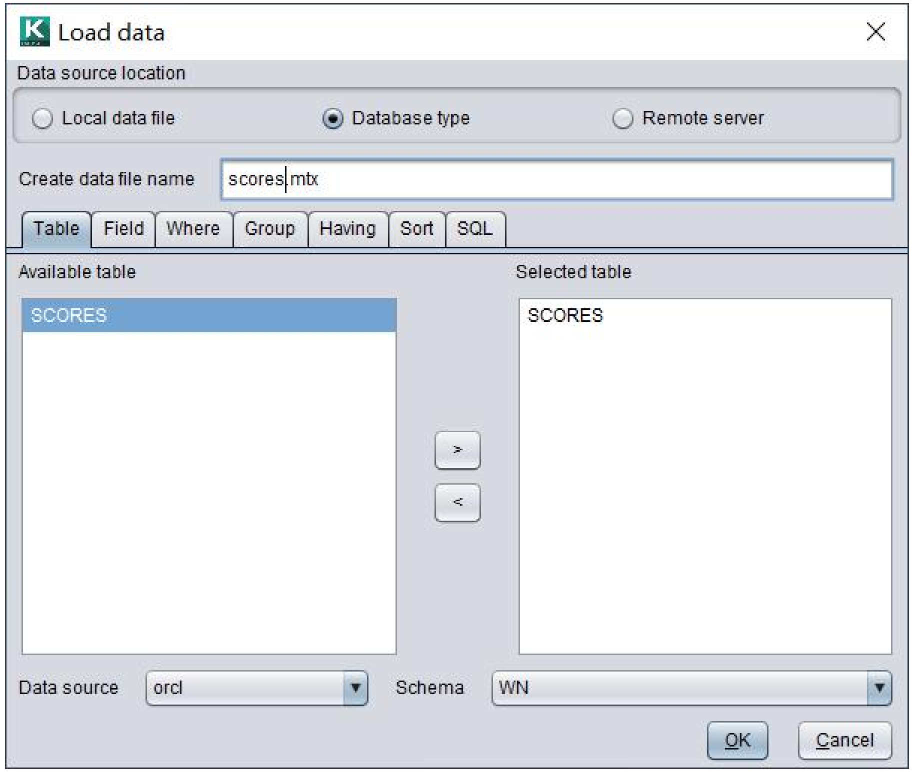


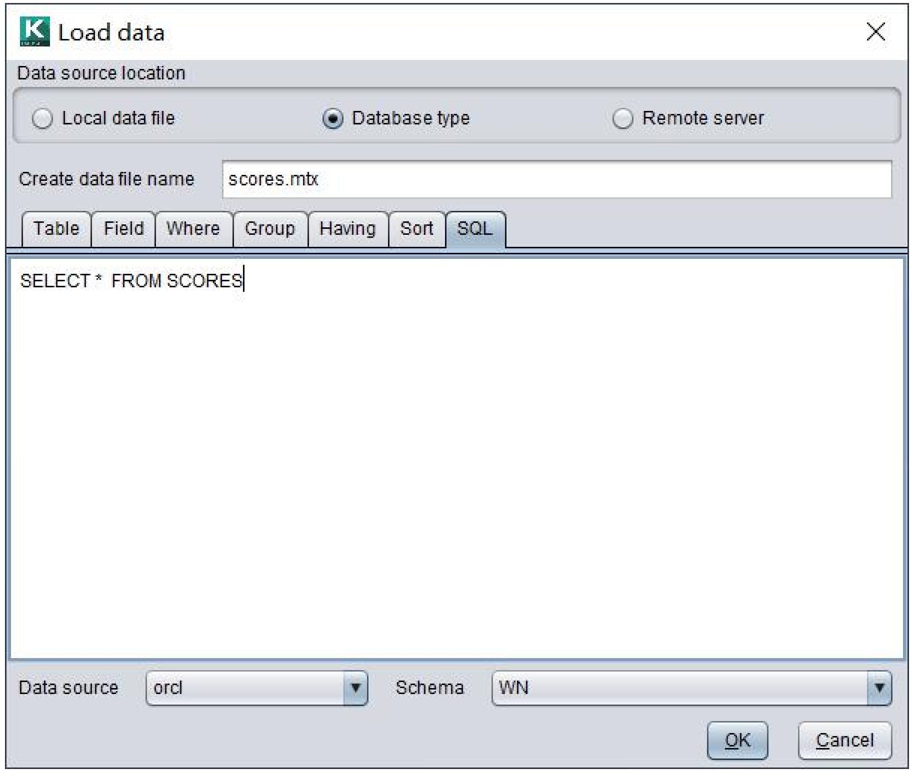
ODBC Datasource



2. Database

Next, you can use the configured data source to edit the SQL statement for data loading.





CONTENTS

02

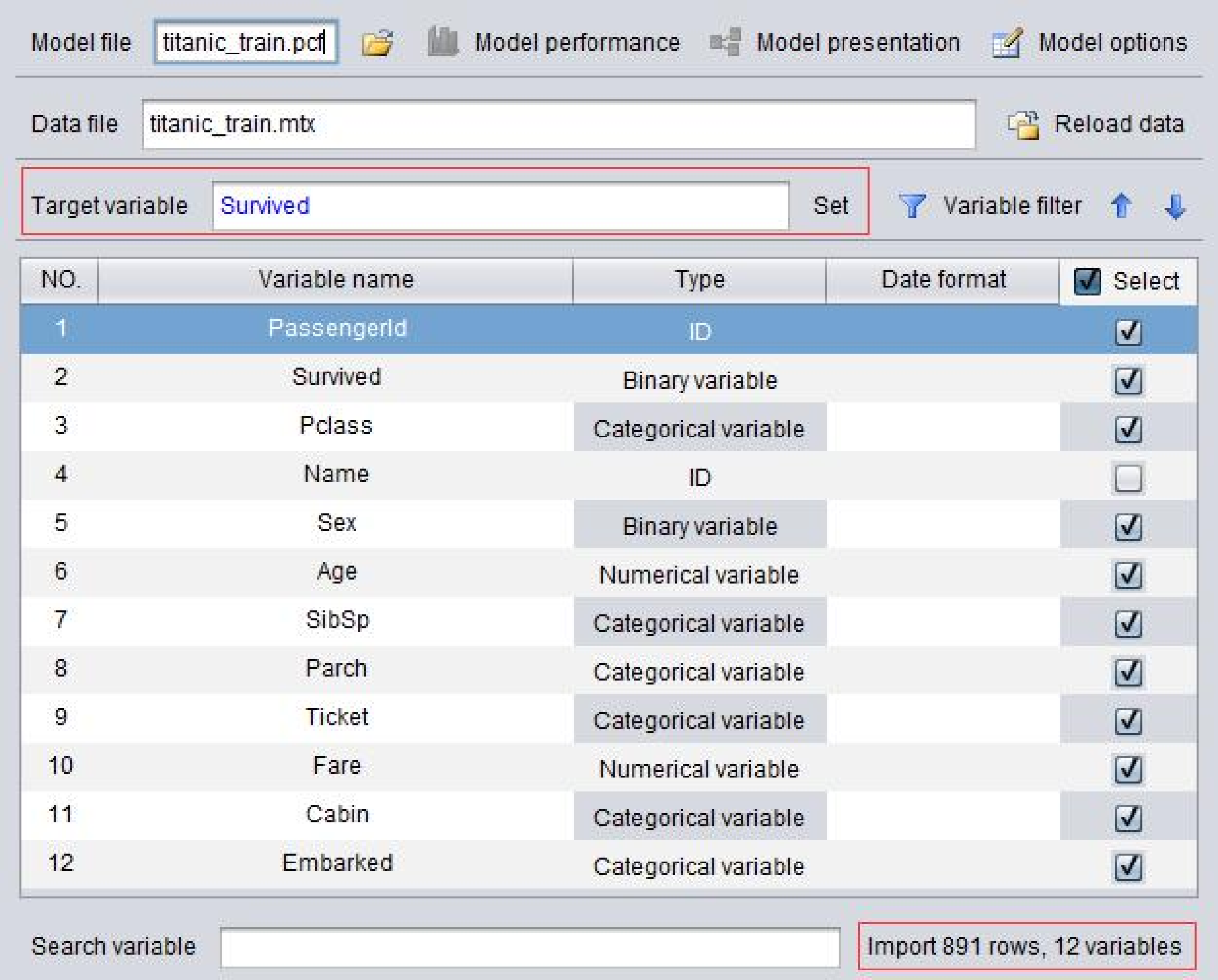
Data Exploration

1. Basic characteristics

After importing the data, the basic characteristics of the data are displayed:

The target variable is survived (it needs to be set by the user), with 12 variables and 891 records.

Automatically parses the types of each variable and the recommended selection status.



1. Basic characteristics

The variable types of intelligent modeling are as follows:

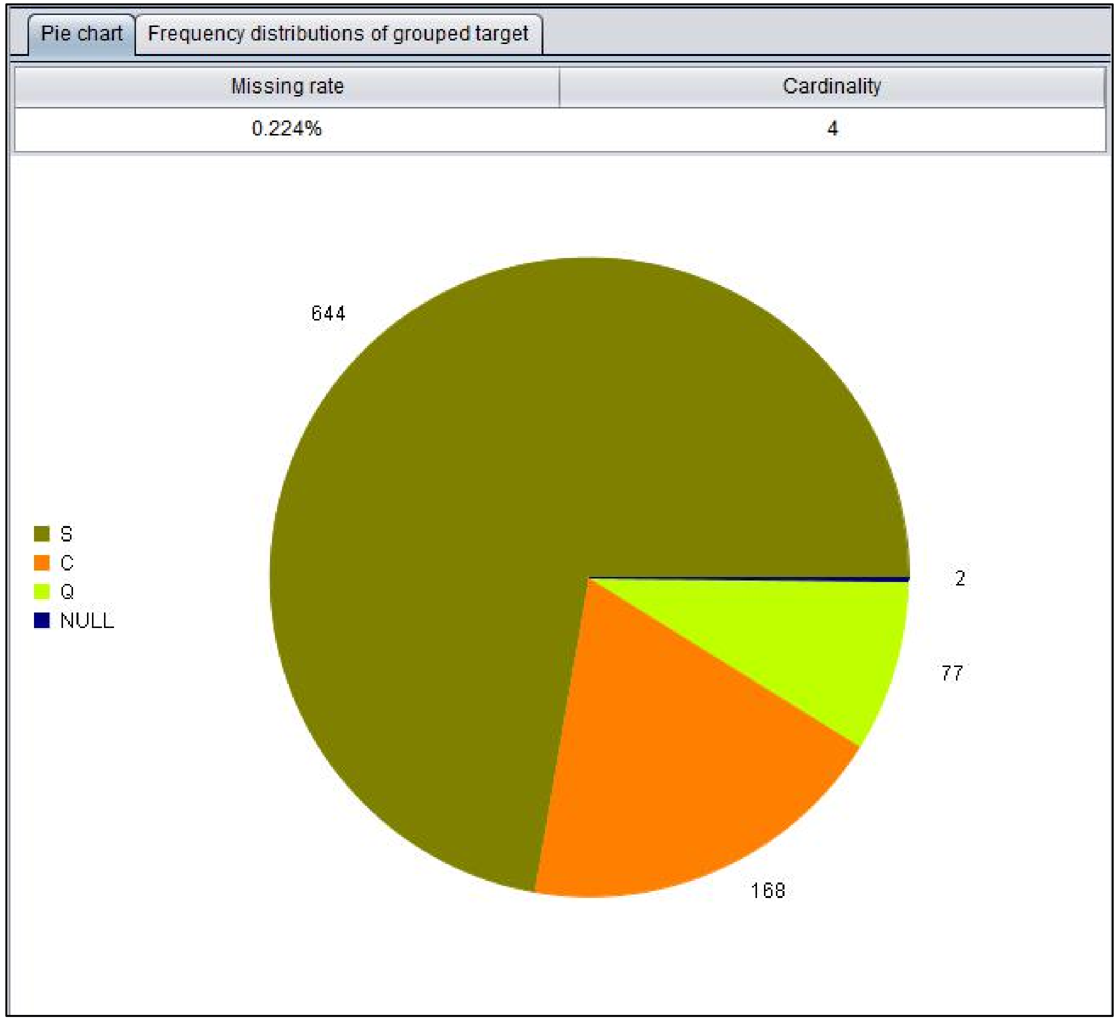
Variable type	Description
Numerical variable	Variable with real number value
Single value variable	Variables containing only one category (excluding missing values)
Binary variable	Variables with only two categories (excluding missing values)
Count variable	Variable with natural value
Categorical variable	Variables with more than two classifications (excluding missing values)
ID	Unique identifier
Time and date	Date, time or datetime variable
Long text	Variables with a length of more than 128 bytes and a large number of classifications

The target variables of intelligent modeling support binary variables, numerical variables, count variables and categorical variables.

2. Statistics of discrete variables

Discrete variables include single value variables, binary variables and categorical variables.

Missing rate: the percentage of missing values in all data.
Potential: the number of members of the set that can be valued by a discrete variable.
Pie chart shows the proportion of each classification.



2. Statistics of discrete variables

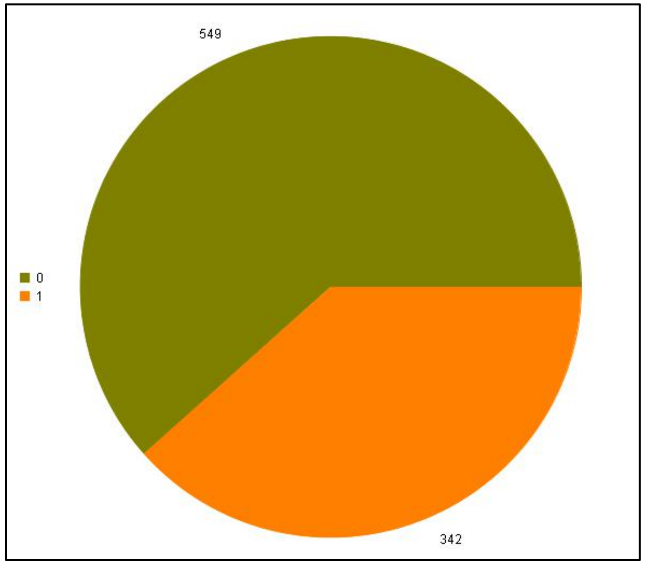
Target variable is binary variable: frequency distribution table of grouped target

In the frequency distribution table of grouped target, samples are grouped according to the classification value, and the number of samples in each group, the number of positive samples and the rate of positive samples are observed.

The positive sample of binary target variable refers to the classification value with a small number of samples. As can be seen from the right figure, in this example, the positive sample is a record with a target variable value of 1.

Pie chart | Frequency distributions of grouped target

Categorical variable	Sample size	Positive cases size	Positive cases rate
S	644	217	33.696%
C	168	93	55.357%
Q	77	30	38.961%
NULL	2	2	100%



Pie chart of target variable

2. Statistics of discrete variables

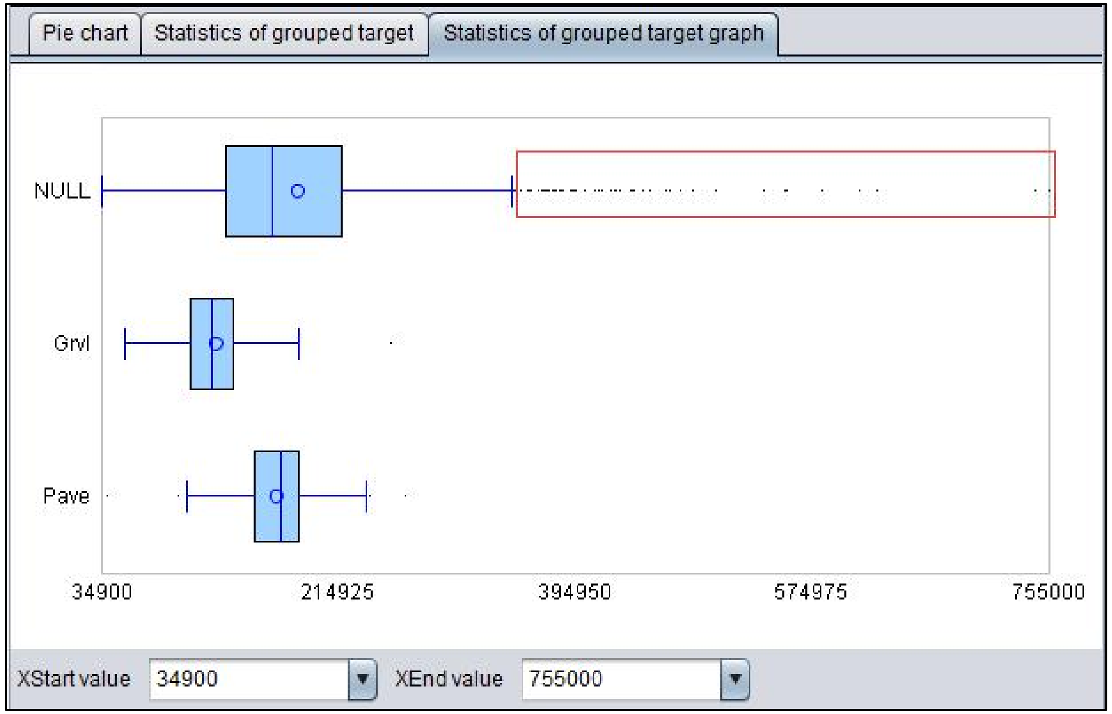
Target variable is numerical variable: statistics of grouped target, statistics of grouped target graph

Grouped target statistics group the samples according to the categorical value, and observe the statistics of each group of samples. Including: frequency, average, standard deviation, median, minimum and maximum.

Pie chart | Statistics of grouped target | Statistics of grouped target graph

Categorical variable	Frequency	Average	Standard deviation	Median	Minimum	Maximum
NULL	1369	183452.131	80667.145	165000	34900	755000
Grvl	50	122219.08	34780.781	119000	52500	256000
Pave	41	168000.585	38370.375	171900	40000	265979

The statistical graph of grouped target, in the form of box line chart, more intuitively represents the distribution of each group of samples. A box line chart can be used to mark outliers.



3. Continuous variable statistics

Continuous variables include numerical variables, count variables and time date variables.

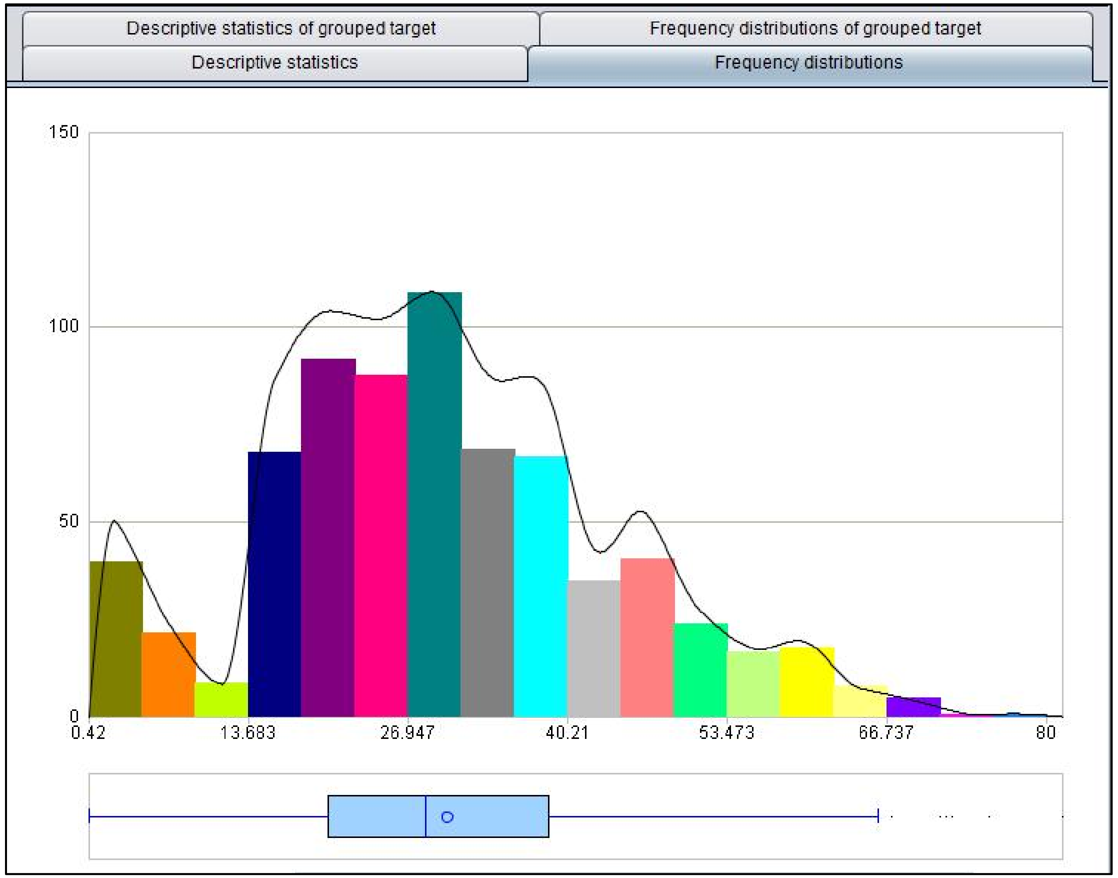
Descriptive statistics show the basic statistical information of the data.

Frequency distribution diagram includes frequency distribution histogram, normal distribution curve and box line chart.

Descriptive statistics of grouped target | Frequency distributions of grouped target

Descriptive statistics | Frequency distributions

Missing rate	Minimum	Maximum	Average	Upper quartile	Median	Lower quartile	Standard deviation	Skewness
19.865%	0.42	80.0	29.699	38.0	28.0	20.0	14.526	0.388



3. Continuous variable statistics

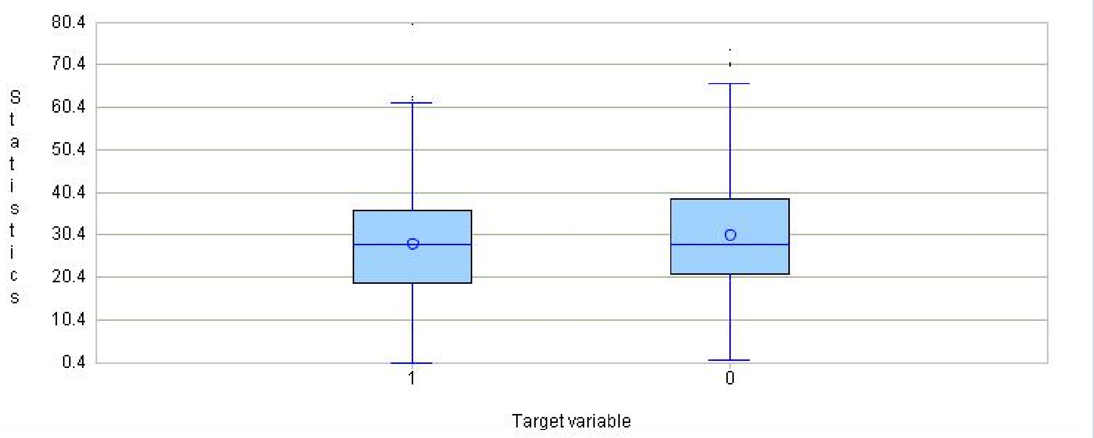
Target variable is binary variable: descriptive statistics of grouped target

Descriptive statistics of grouped target group the samples according to the target variable values, make statistics respectively, and draw the corresponding box line chart.

Descriptive statistics | Frequency distributions

Descriptive statistics of grouped target | Frequency distributions of grouped target

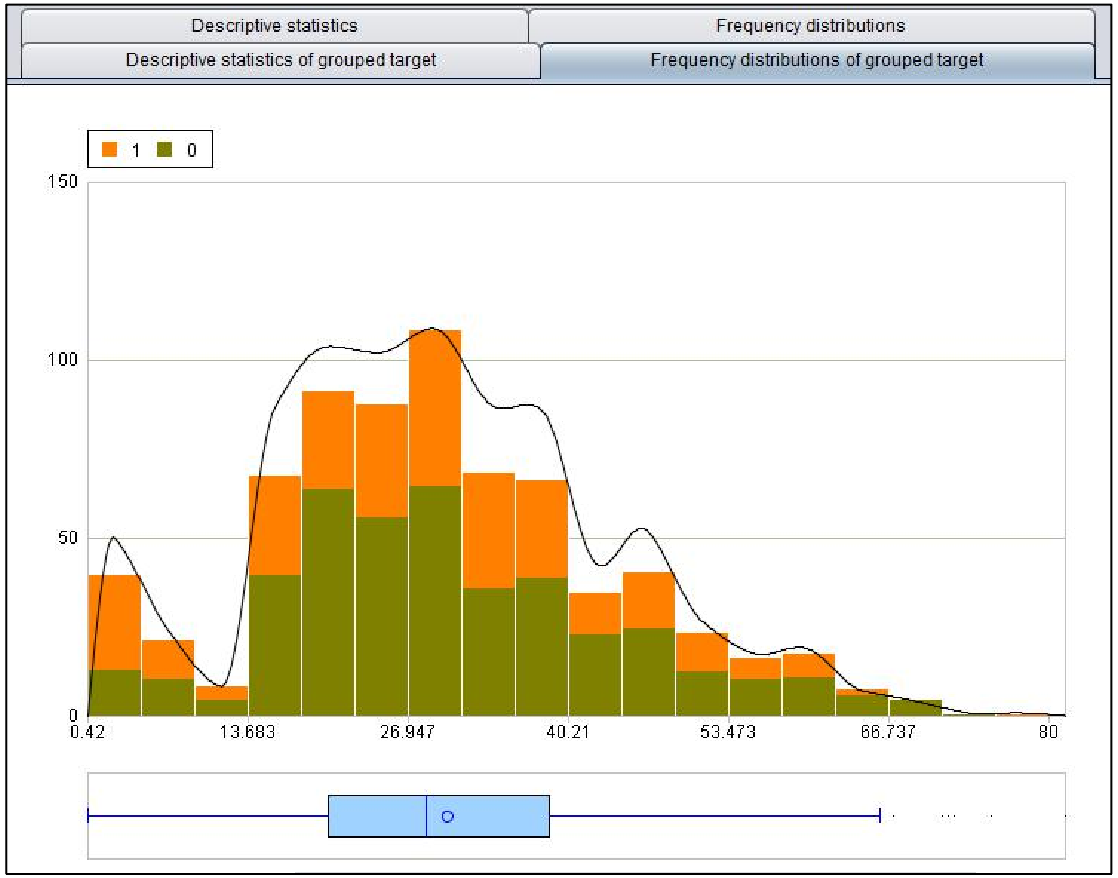
	Target=1	Target=0	Total
Missing rate	15.205%	22.769%	19.865%
Average	28.344	30.626	29.699
Standard deviation	14.951	14.172	14.526
Skewness	0.18	0.584	0.388
Minimum	0.42	1.0	0.42
Lower quartile	19.0	21.0	20.0
Median	28.0	28.0	28.0
Upper quartile	36.0	39.0	38.0
Maximum	80.0	74.0	80.0



3. Continuous variable statistics

Target variable is binary variable： frequency distributions of grouped target

Frequency distributions of grouped target: the samples in each interval are grouped according to the target variable value, and the frequency is displayed in different colors.



3. Continuous variable statistics

Target variable is a numerical variable: target variable correlation coefficient

Pearson correlation coefficient: used to describe the linear correlation between two continuous variables.

Spearman rank correlation coefficient: used to describe the rank correlation between two continuous variables.

The greater the absolute value of the correlation coefficient, the greater the correlation between the two variables.

Descriptive statistics | Frequency distributions
Target variable correlation coefficient | Single factor scatter plot

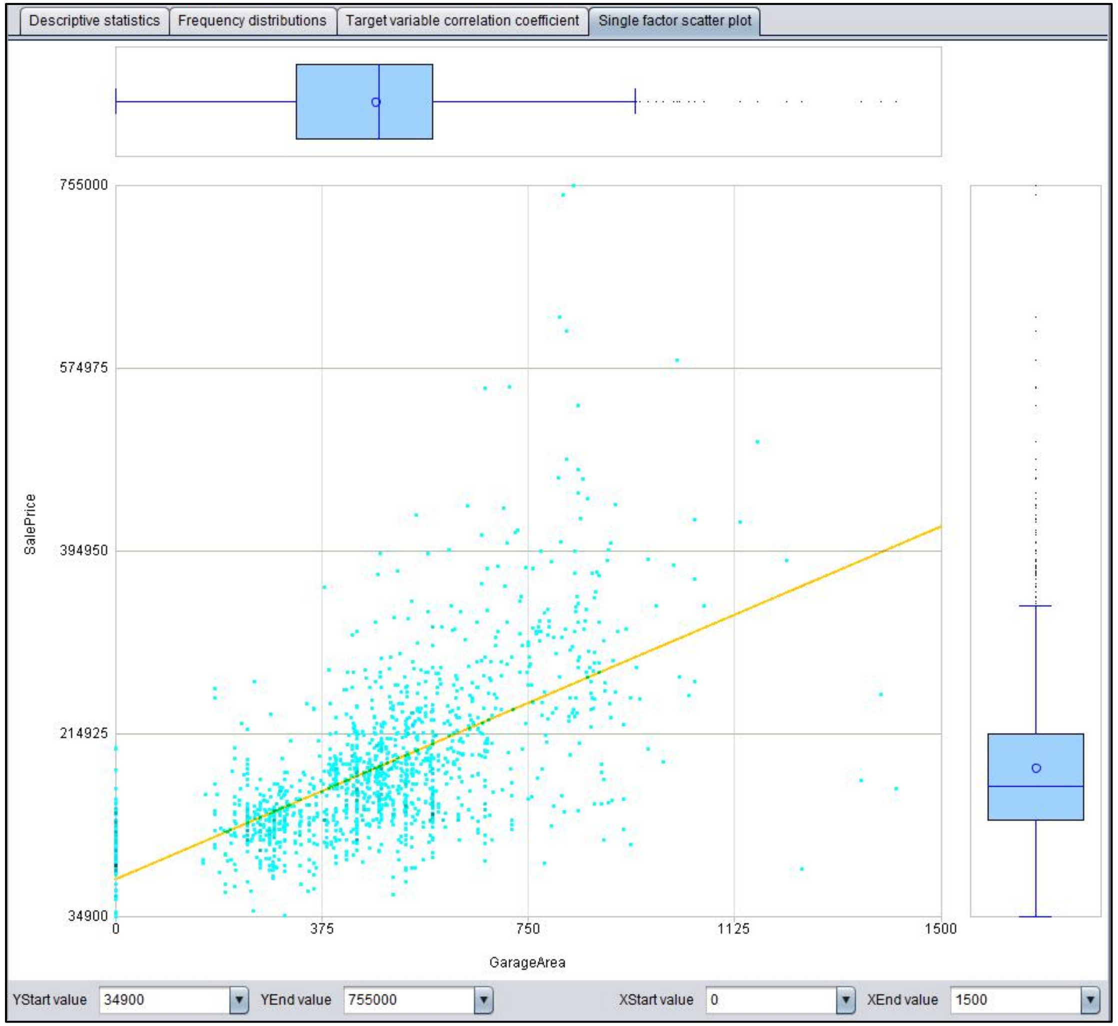
Pearson	Spearman
0.6234	0.6494

Above is the correlation coefficient between garage area and house price. It can be seen that there is a strong correlation between the two.

3. Continuous variable statistics

Target variable is a numerical variable： single factor scatter plot

The single factor scatter plot intuitively shows the correlation distribution of current variable (garage area) and target variable (house price). The yellow line is the regression line.



4. Data quality report

Provide the function of exporting data quality report to PDF file. Some contents are as follows:

These observations from 891 unique ID. Since the number of unique Id is equal to the number of observations, time sensitive information can not be studied using this data set.

Variables that have all “empty“ value are not exist.

Variables that have more than 99% of missing values are not exist.

Variables that have missing values between 95% and 99% are not exist.

Table 1 Missingness Analysis

Missing Percentage	Number of Variables	% of All Numerical Variables
100%	0	0%
99% to 100%	0	0%
95% to 99%	0	0%
90% to 95%	0	0%
80% to 90%	0	0%
70% to 80%	0	0%
60% to 70%	0	0%
50% to 60%	0	0%
30% to 50%	0	0%
10% to 30%	1	20%
Below10%	4	80%

The highly positive skewness (with skewness > 10) numerical variables are not exist.

The highly negative skewness (with skewness < -10) numerical variables are not exist.

Table 2 Skewness of Numerical Variables

Skewness Range	Number of Variables	% of All Numerical Variables
10+	0	0%
5 to 10	0	0%
2 to 5	3	60%
1 to 2	0	0%
-1 to 1	2	40%
-2 to -1	0	0%
-5 to -2	0	0%
-10 to -5	0	0%
-10-	0	0%
Total	5	100%

All categorical variables with cardinality over 512 are Name,Ticket.

The calculation of cardinality includes missing category.

The following categorical variables have cell frequency less than 100: Name,TicketSurvived,Pclass,Sex,Embarked.

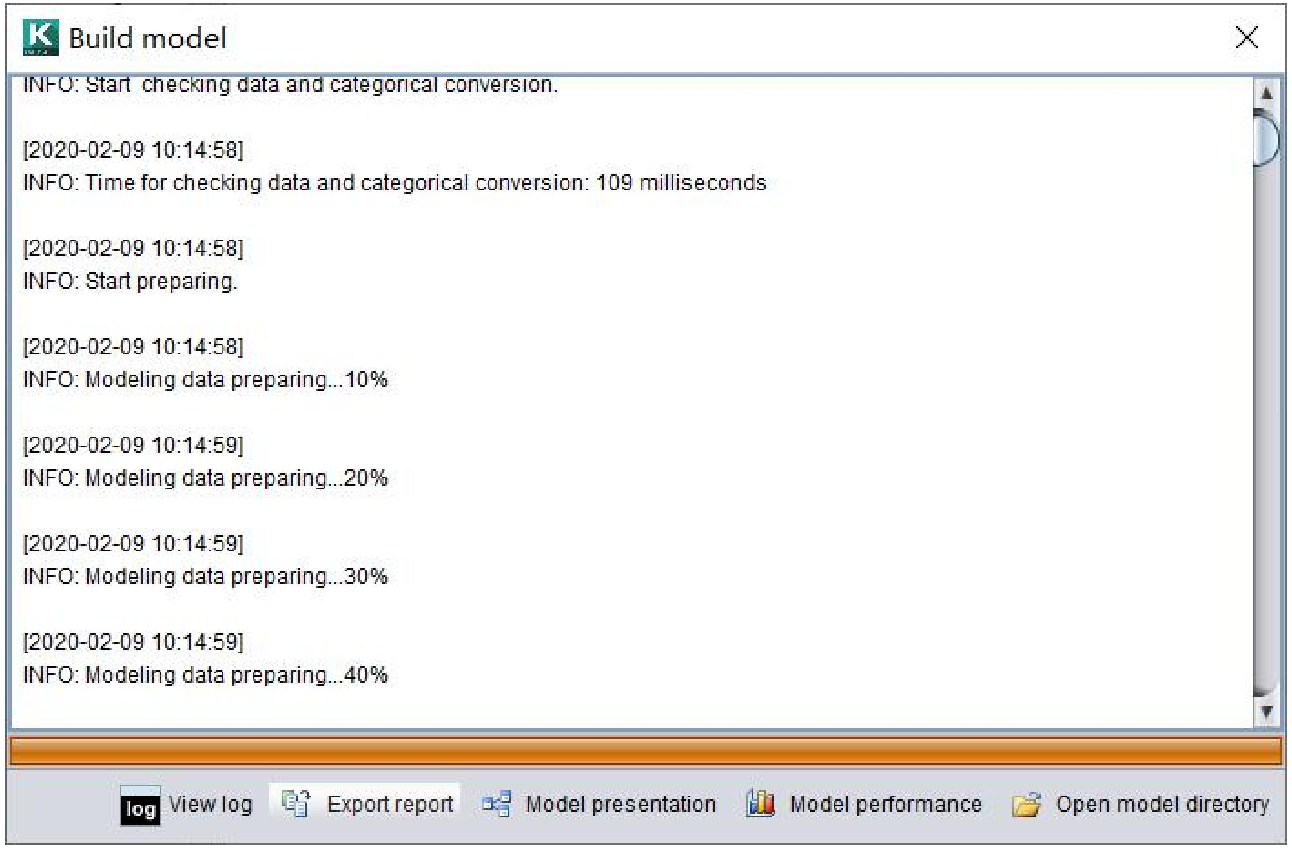
CONTENTS

03

Preprocessing

1. Automatic preprocessing

The preprocessing process of intelligent modeling is integrated in the modeling process, with one key automatic preprocessing.



2. Preprocessing report

After modeling, you can export the model report, which describes the actions of preprocessing. Some contents are as follows:

Target variable: Survived, ID variable: PassengerId.
The number of fields before pretreatment: 12, the number of fields after pretreatment: 11.
The number of fields with missing values before pretreatment: 3 and the number of fields with missing values after pretreatment: 0.
Total rows of data: 891, where deleted rows due to missing target: 0.

Variable selection table

	Number of selections	Number not selected	Total number
All variables	11	1	12
Unary variables	0	0	0
Binary variables	2	0	2
Category variables	4	1	5
Numerical variables	2	0	2
Counting variables	2	0	2
Datetime variables	0	0	0

Variables Processing Information

Variable name: PassengerId. The type is ID
Variable name: Pclass. The type is Category variables

Number of categories: 3
The variable fills the missing value by using the yimming intelligent filling algorithm.
There are 3 categories are merged because of low frequency.
Generation Category Derivative Variables: BI_Pclass_1, BI_Pclass_2
Variable name: Sex. The type is Binary variables
Number of categories: 2
The variable fills the missing value by using the yimming intelligent filling algorithm.
There are 2 categories are merged because of low frequency.
Generation Category Derivative Variables: BI_Sex_1
Variable name: Age. The type is Numerical variables
Skewness: 0 Average:29.699

Median:24 Variance:13.002
The variable fills the missing value by using the yimming intelligent filling algorithm.
Variable name: SibSp. The type is Counting variables
Skewness: 0 Average:0.523
Median:0 Variance:1.103
The variable fills the missing value by using the yimming intelligent filling algorithm.
Variable name: Parch. The type is Counting variables

3. Preprocessing process

(1) Check variable value field

Check and record the value range of all variables. If the test data has a category that is not listed in the training data or beyond the range of values, certain processing needs to be carried out.

(2) Time date variable processing

Check all time and date variables and create several commonly used derived variables. Check the correlation of time and date variables, and create multi date linkage derived variables.

(3) Missing value information extraction

If there are missing values in the data, the missing value pattern is extracted and recorded, and the behavior characteristics of missing values are transformed into derivative variables for use.

3. Preprocessing process

(4) Missing value filling

If there are missing values in the data, use simple or personalized intelligent algorithm to fill in the missing values.

(5) Noise reduction of categorical variables

For the noise that may exist in the categorical variables, such as very few category, abnormal category, suspected error classification and so on, carry out targeted processing.

(6) Convert the categorical variable to a numeric variable

Convert the categorical variable to a numeric variable that can be operated normally. The main method is dummy variable and smoothing, which is judged by algorithm intelligence.

3. Preprocessing process

(7) Rectify deviation

For some models with normal hypothesis, the high skewness variables are transformed mathematically to make the skewness return to 0, which satisfies the model hypothesis.

(8) Exception handling

Detect and identify possible outliers, and deal with them accordingly.

(9) Variable selection

In order to reduce the time cost and the complexity of the model, we need to remove the useless variables.

3. Preprocessing process

(10) Standardization / normalization

Data standardization / normalization to eliminate caliber difference. It is beneficial to the optimization of neural networks and other models.

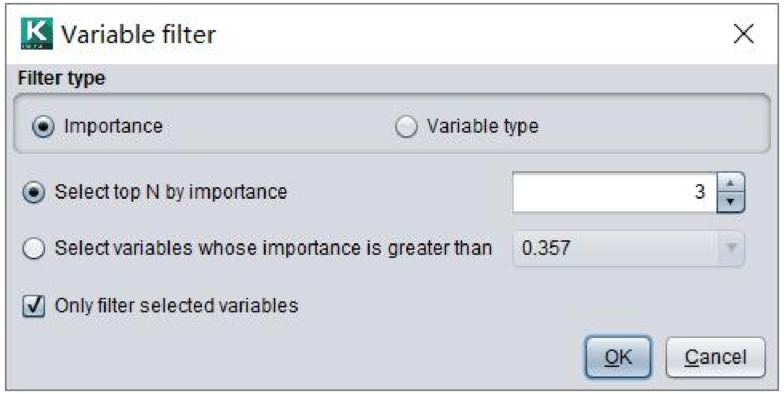
(11) Sample balancing

For binary data, if the proportion of positive and negative samples is seriously unbalanced, it will be balanced according to the specified proportion, and intelligent resampling modeling will be carried out.

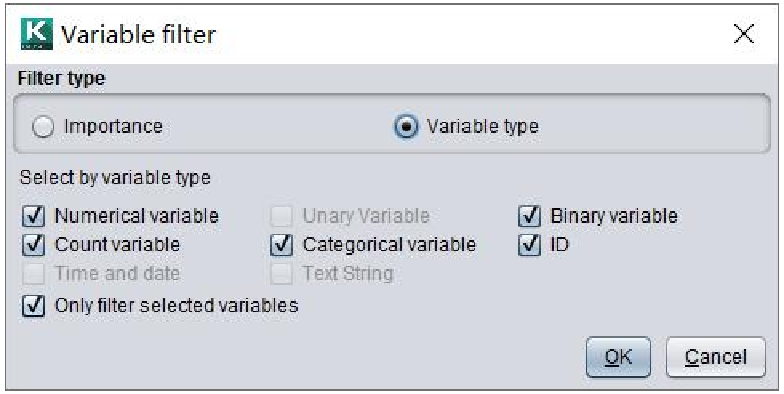
4. Manual preprocessing

Variable selection

Remove some irrelevant variables according to the variable type. For example, ID and long text, single value variable without missing value, etc.



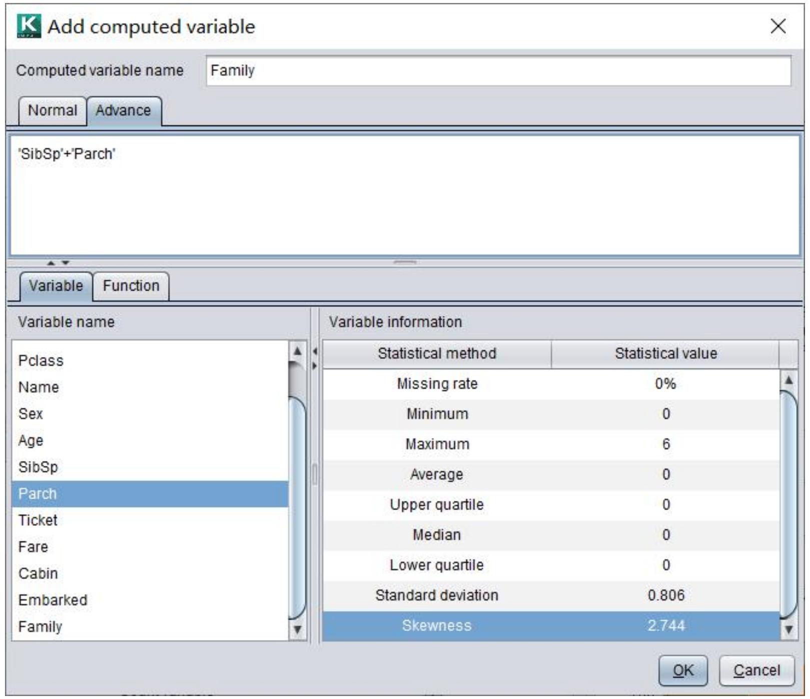
Filter variables according to the importance of variables, only the variables with higher importance are retained. Variable importance can be imported from data dictionary or obtained through modeling.



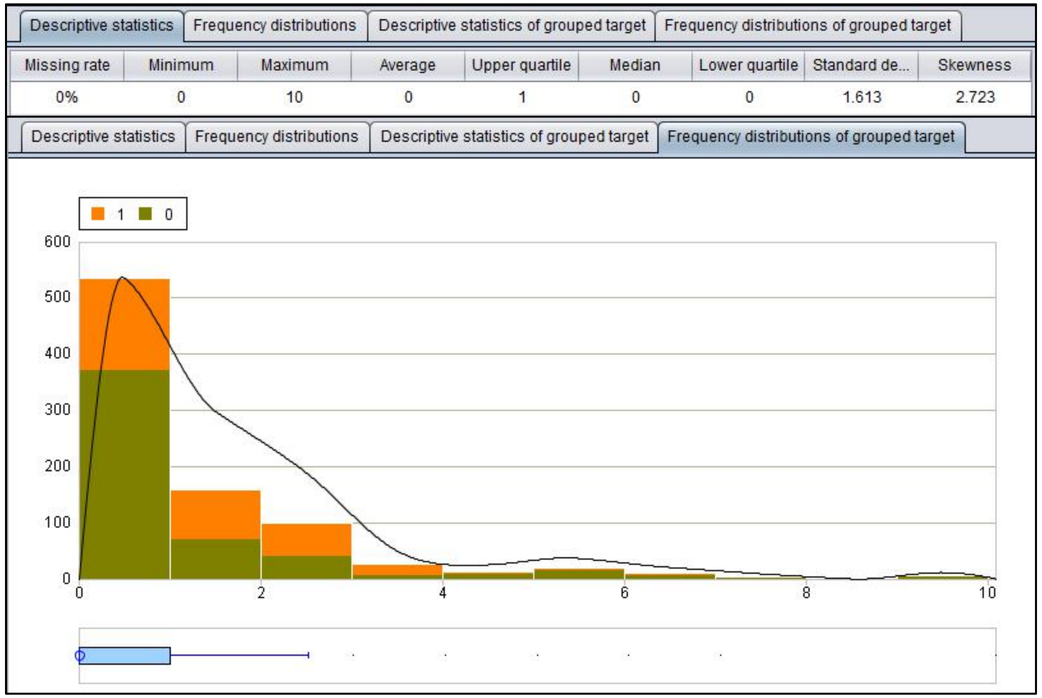
4. Manual preprocessing

Derived variables

The number of family members is obtained by adding the number of variable “SibSp" and the number of variable “Parch". It can be seen that the survival rate of family members is higher at 1-3.



Add derived variable family

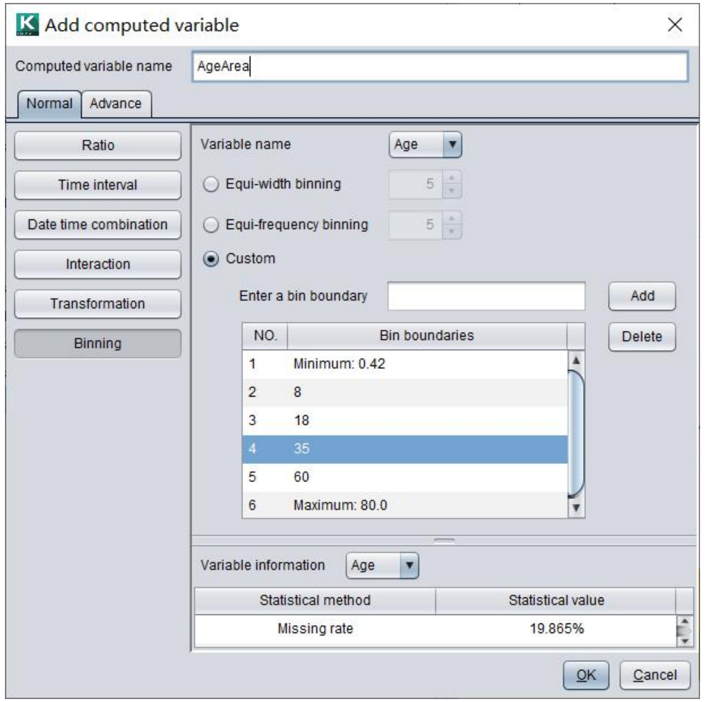


Variable family statistics

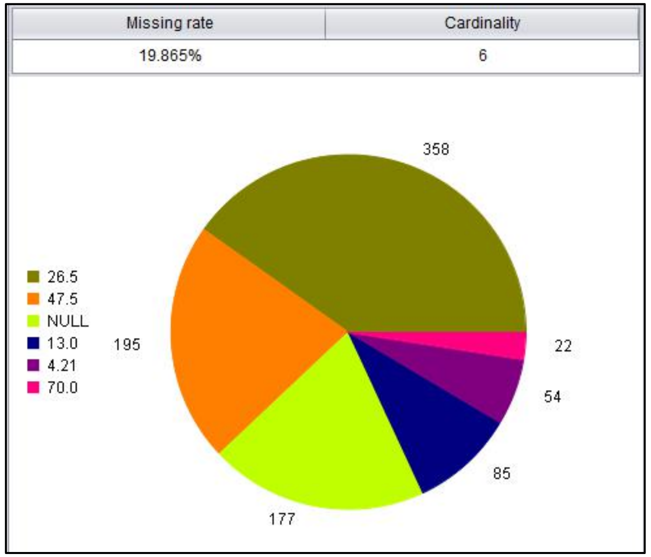
4. Manual preprocessing

Derived variables

The numerical variables can be discretized and converted into categorical variables. Taking age as an example, it is divided into 0, 8, 18, 35 and 60 age groups, generating derivative variables and making statistics.



Add derived variable AgeArea



Pie chart | Frequency distributions of grouped target

Categorical variable	Sample size	Positive cases size	Positive cases rate
NULL	177	52	29.379%
4.21	54	36	66.667%
13.0	85	34	40%
26.5	358	137	38.268%
47.5	195	78	40%
70.0	22	5	22.727%

Variable AgeArea statistics

It can be seen that the survival rate of the 0-8-year-old is the highest, the difference between the young and the middle-aged is not big, and the survival rate of the old is the lowest.

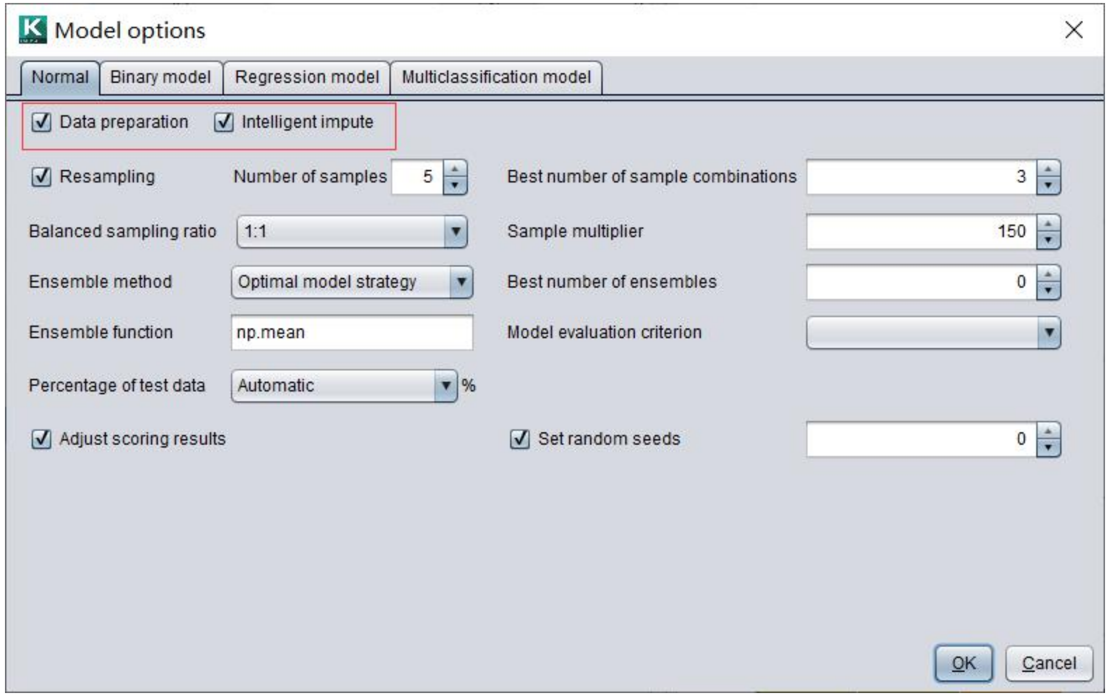
4. Manual preprocessing

Preprocessing options

In the model options, you can define whether to preprocess data and whether to fill it intelligently.

If the data has been preprocessed, you can cancel the data preprocessing.

Intelligent filling can better fill the missing value, but it will consume more hardware resources and time. When the amount of data is large, intelligent filling is not recommended. If unchecked, it will be filled in simply.



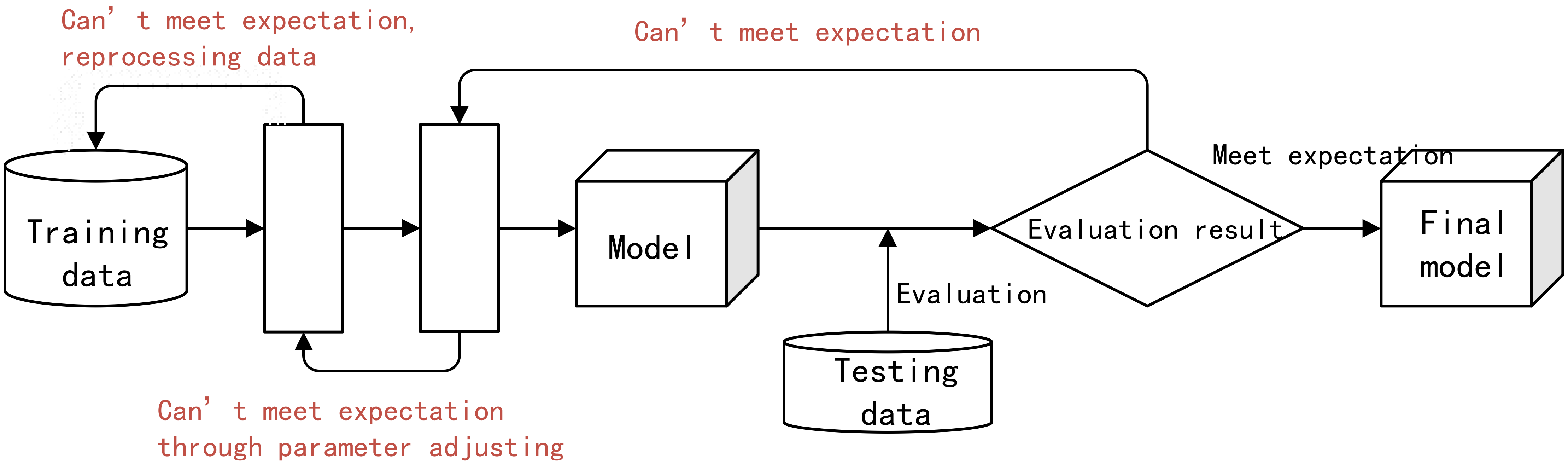
CONTENTS

04

Modeling

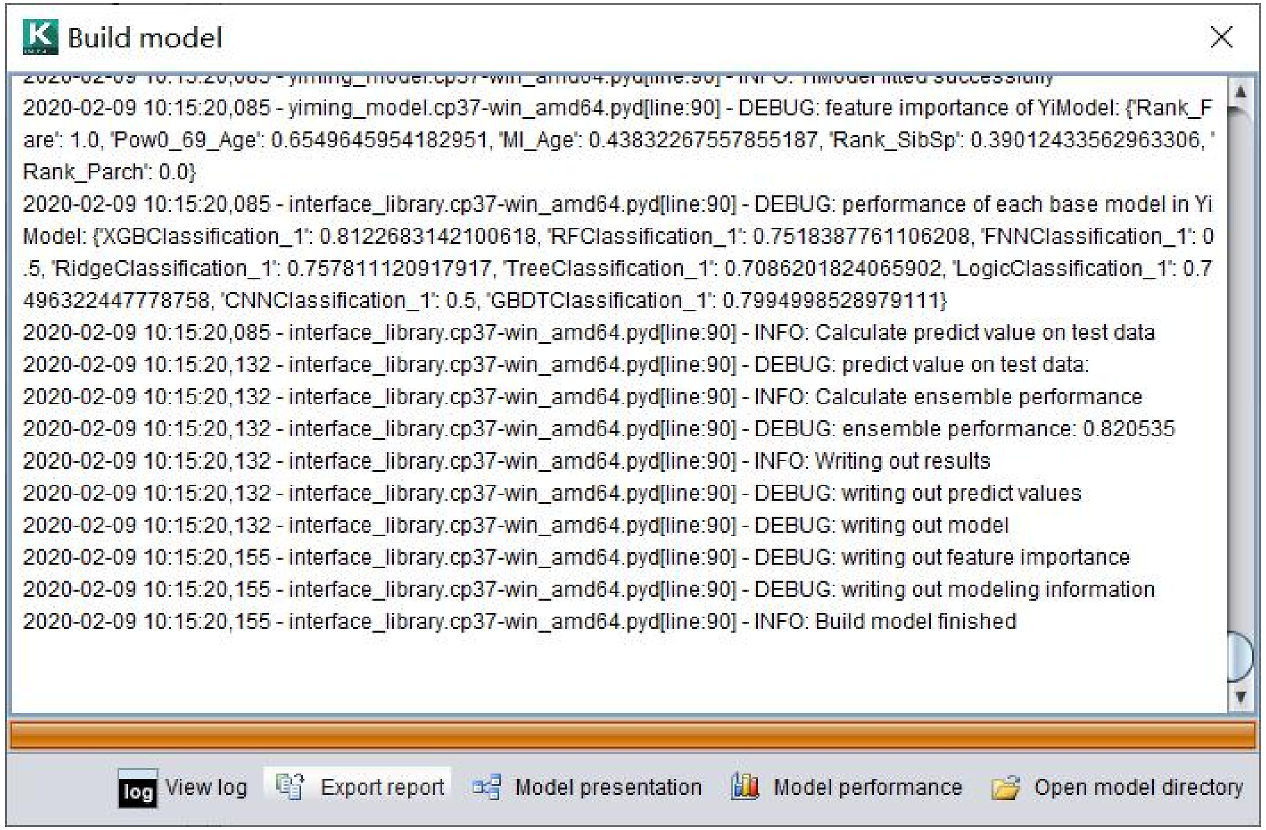
1. Modeling process

When using traditional tools, it usually requires professionals with statistical basis to continuously select algorithms, adjust model parameters, and finally get the expected model. The modeling process is as follows:



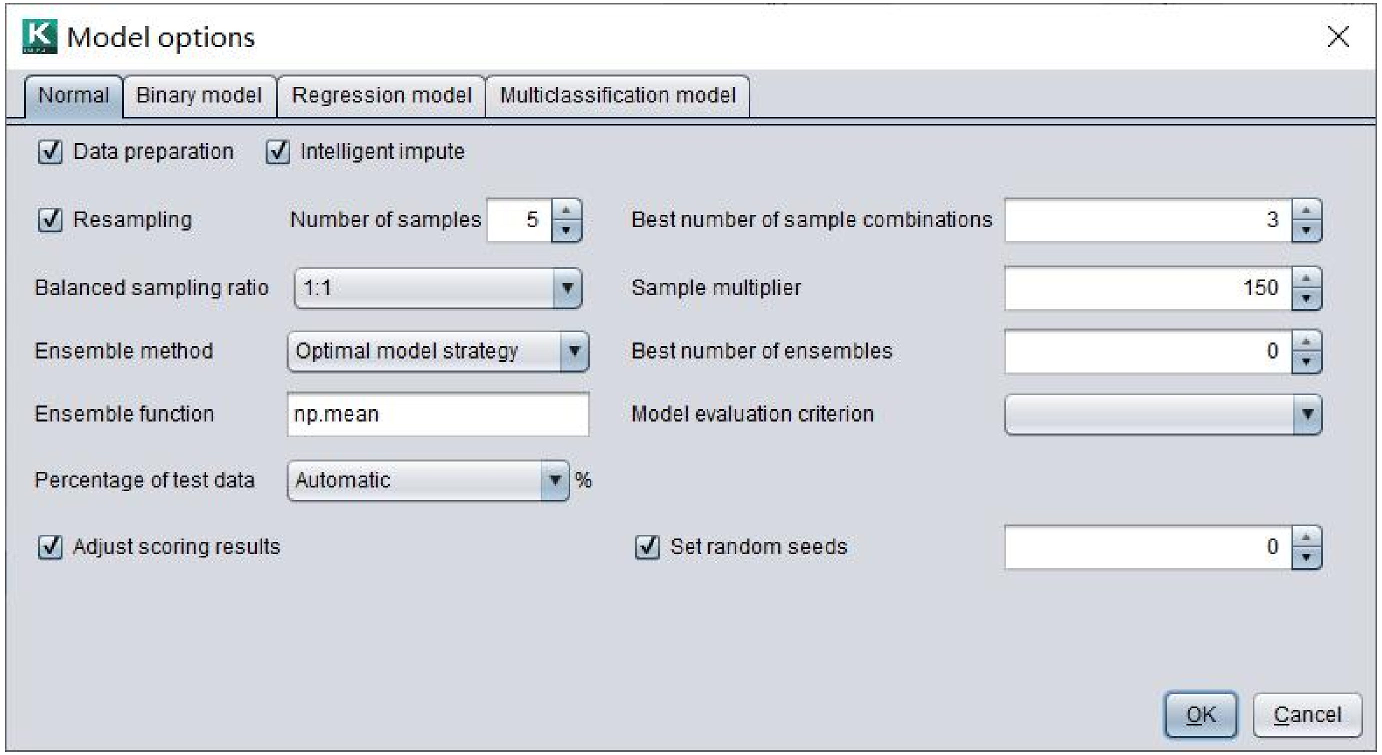
2. Intelligent modeling

Intelligent modeling tools do not need statistical knowledge, one key intelligent modeling, optimization of model combination and model parameters are implemented internally.



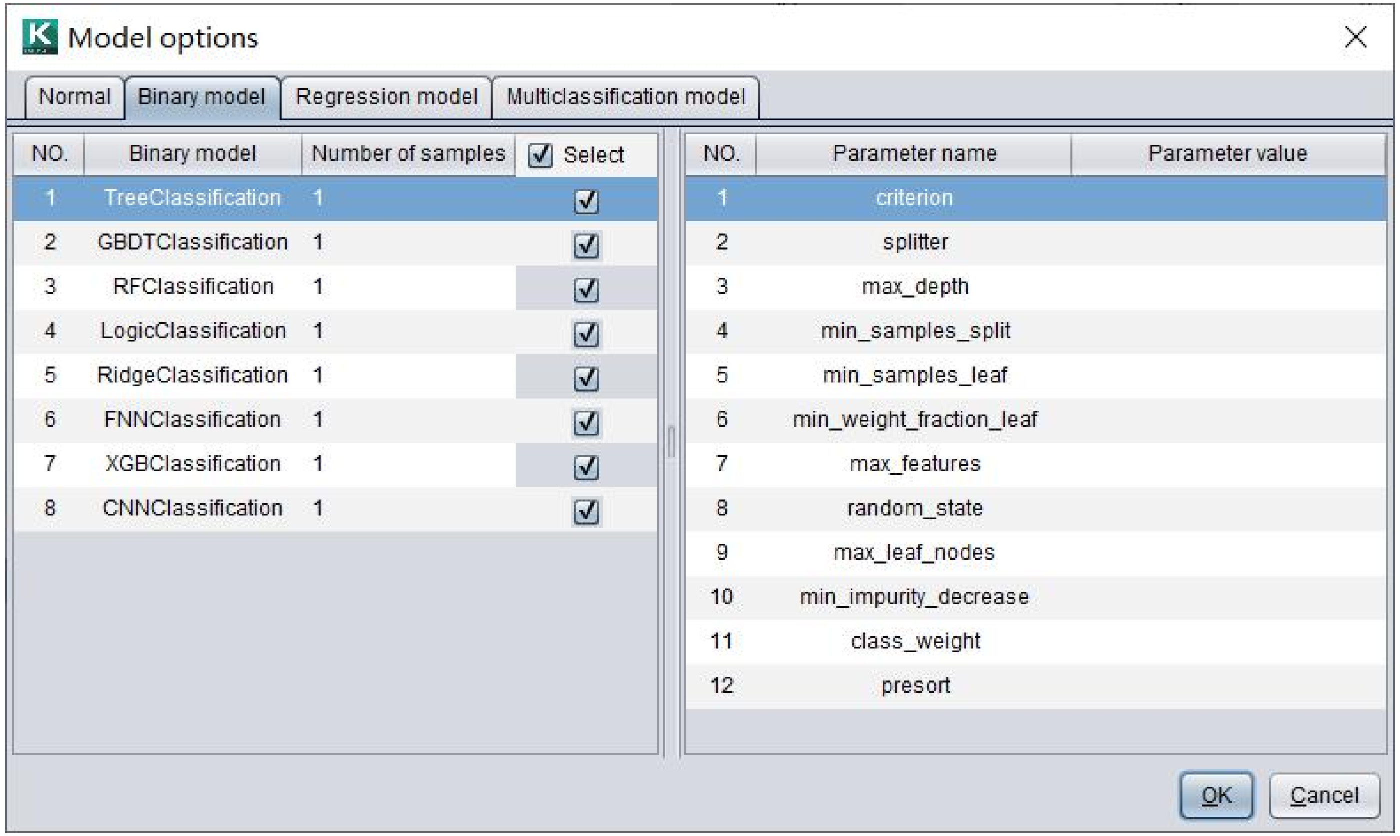
3. Professional modeling

Intelligent modeling opens up model parameters for professional users who are proficient in the models. Here are the general options for the model:



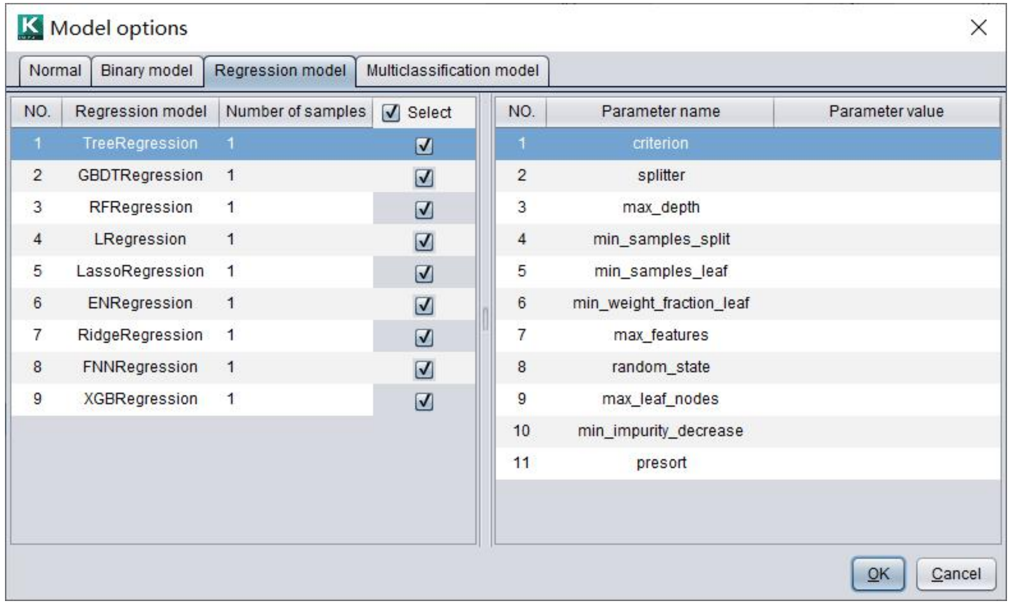
3. Professional modeling

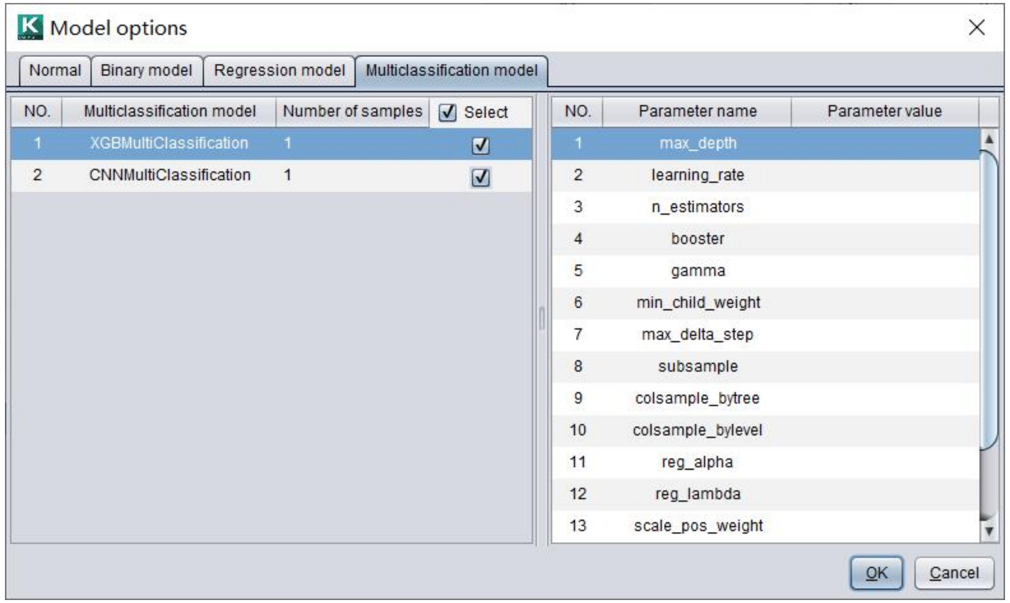
Intelligent modeling supports several binary classification algorithm models in the graph, and can also set whether each model is used and the sampling times. On the right, you can set parameter values for each model. For ordinary users, these settings can be ignored.



3. Professional modeling

Similarly, we can set whether to use regression model and multi classification model, and their respective parameters.





Detailed documentation of each model parameter： http://doc.raqsoft.com/AIModel/userrefer/jm9.html

CONTENTS

05

Model performance

1. Model performance

Classification model: evaluation index

Intelligent modeling provides three commonly used evaluation indexes for classification model:

Model performance

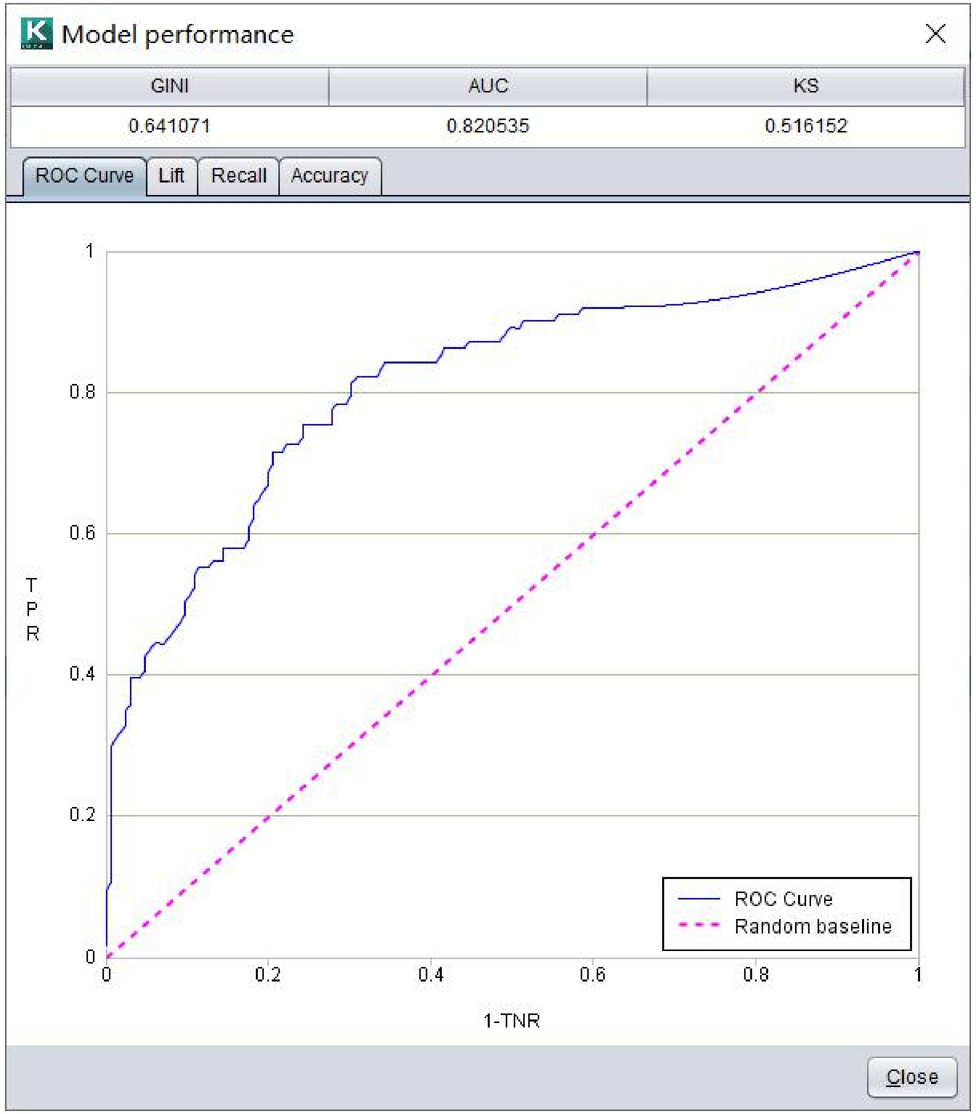
GINI	AUC	KS
0.641071	0.820535	0.516152

Evaluation Index	Description
GINI	Gini index is equal to 2 * auc-1 in numerical value, which is used to characterize the model's ability to distinguish positive and negative samples.
AUC	AUC is equal to the area under ROC curve. The higher AUC is, the better the model is.
KS	KS value is used to measure the ability of the model to distinguish positive and negative samples. The larger the KS value is, the stronger the ability of the model to distinguish positive and negative samples is.

1. Model performance

Classification model: ROC curve

ROC curve is the relationship between true positive class rate and "1-true negative class rate". ROC curve can be regarded as a visual display to evaluate all possible decision-making performance of a given model.

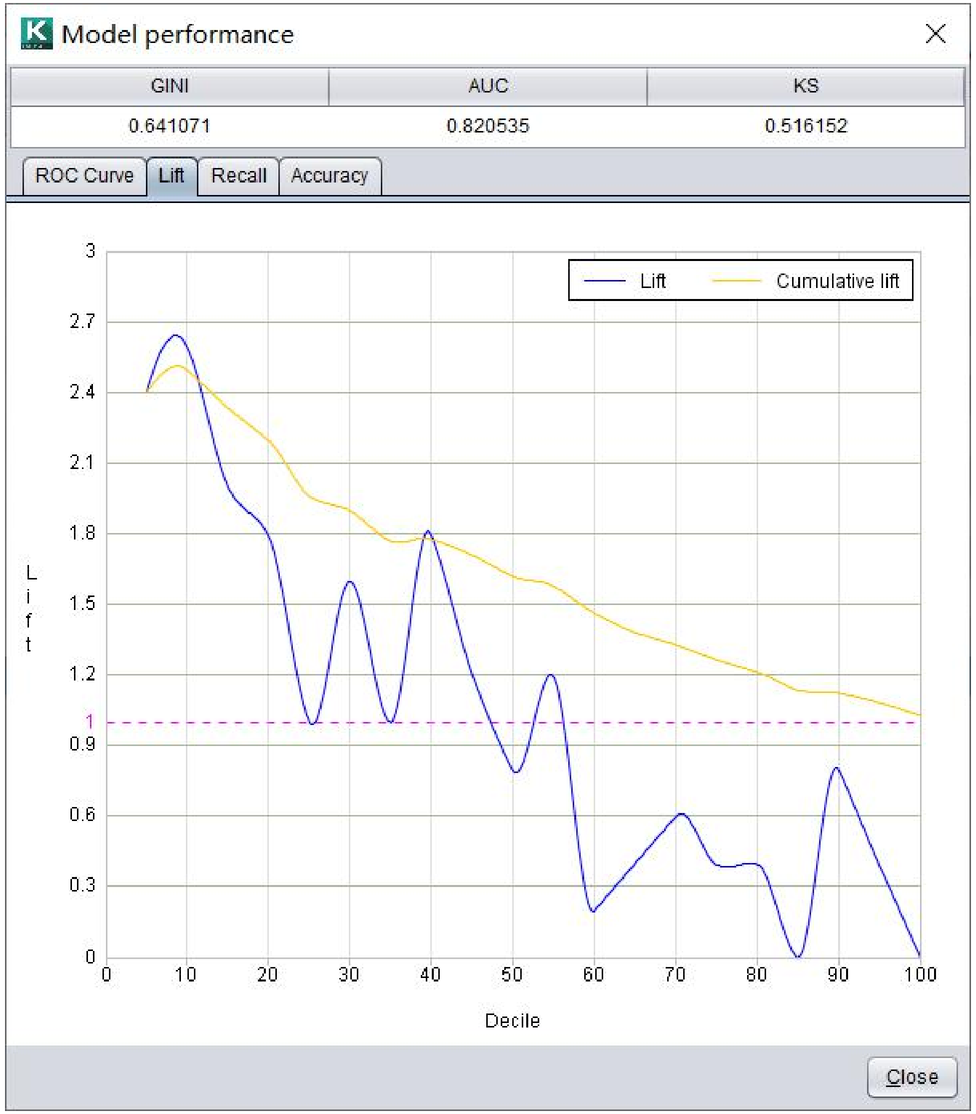


1. Model performance

Classification model: Lift

Lift refers to the multiple that can be improved by using association rules. It is the ratio of the degree of confidence to expected confidence.

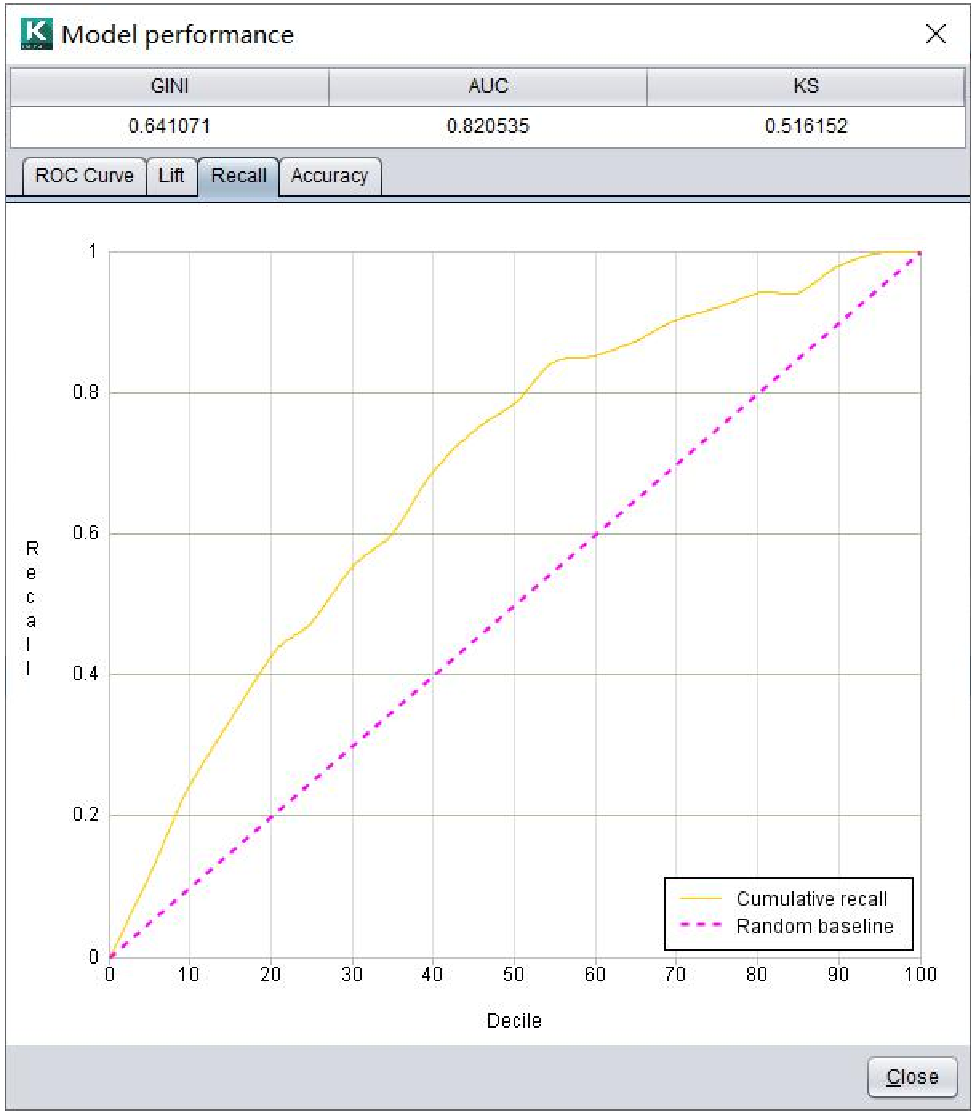
Lift is particularly suitable for targeted marketing and other scenarios.



1. Model performance

Classification model: Recall

Recall graph shows that the model can find positive samples, which is mainly used in the scene of data imbalance. The cumulative recall rate is the ratio of cumulative positive samples and total positive samples in each group.



1. Model performance

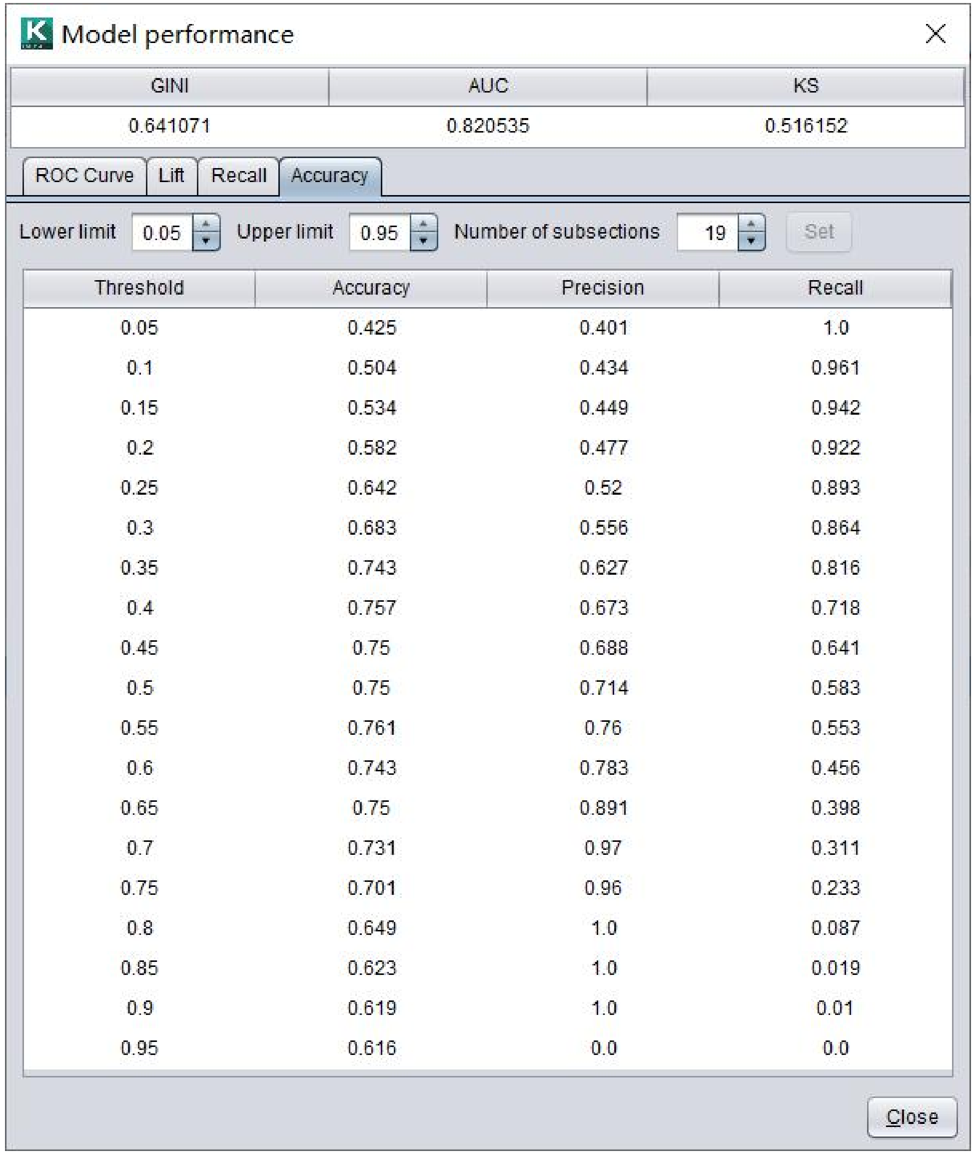
Classification model: Accuracy table

Threshold: value used to distinguish positive and negative samples.

Accuracy: the ratio of correct samples to all samples.

Precision: the correct rate of prediction in the result of positive sample.

Recall: the ratio of correctly predicted positive samples and all positive samples.



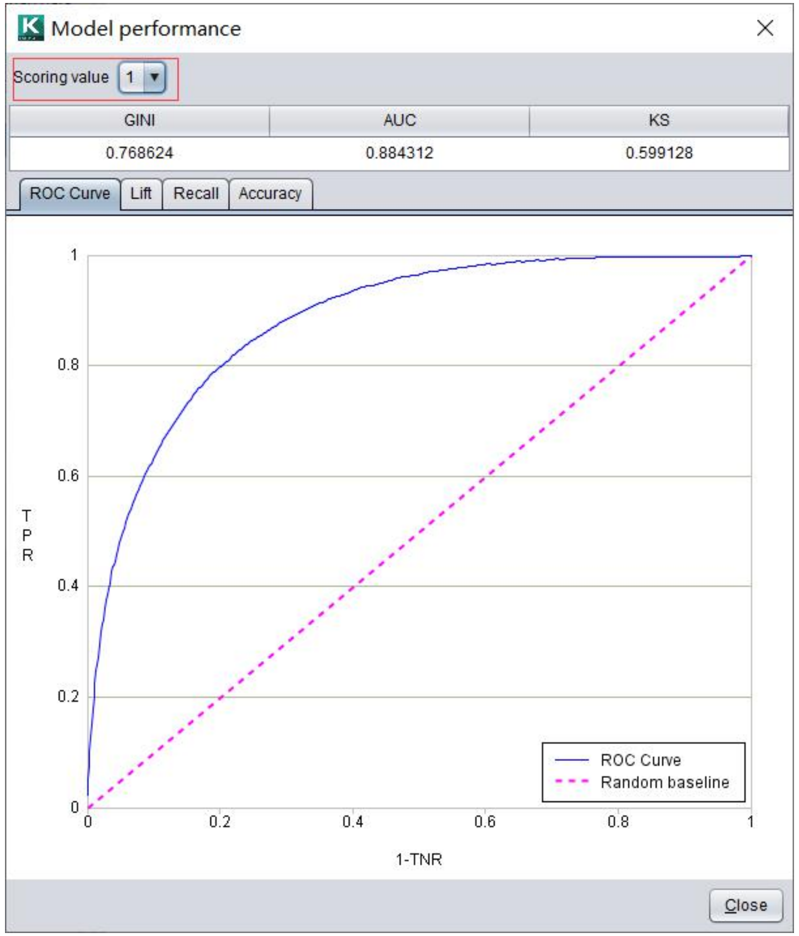
GINI	AUC	KS
0.641071	0.820535	0.516152

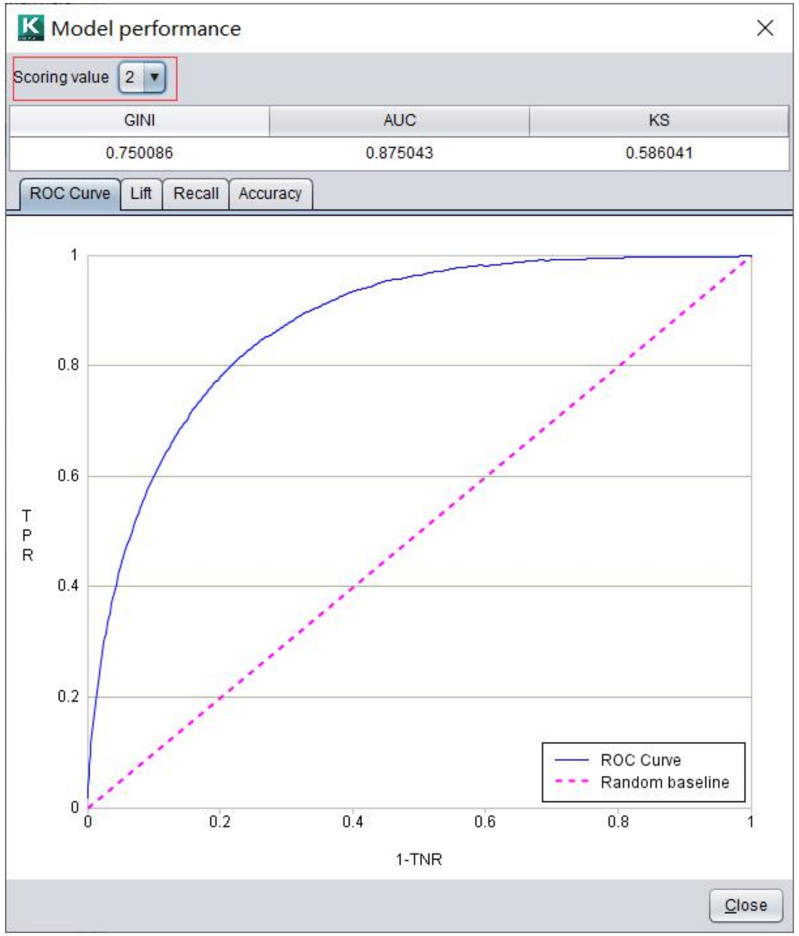
Threshold	Accuracy	Precision	Recall
0.05	0.425	0.401	1.0
0.1	0.504	0.434	0.961
0.15	0.534	0.449	0.942
0.2	0.582	0.477	0.922
0.25	0.642	0.52	0.893
0.3	0.683	0.556	0.864
0.35	0.743	0.627	0.816
0.4	0.757	0.673	0.718
0.45	0.75	0.688	0.641
0.5	0.75	0.714	0.583
0.55	0.761	0.76	0.553
0.6	0.743	0.783	0.456
0.65	0.75	0.891	0.398
0.7	0.731	0.97	0.311
0.75	0.701	0.96	0.233
0.8	0.649	1.0	0.087
0.85	0.623	1.0	0.019
0.9	0.619	1.0	0.01
0.95	0.616	0.0	0.0

1. Model performance

Multiclassification model

When the target variable is a categorical variable, the model performance of each classification can be viewed by switching prediction values.

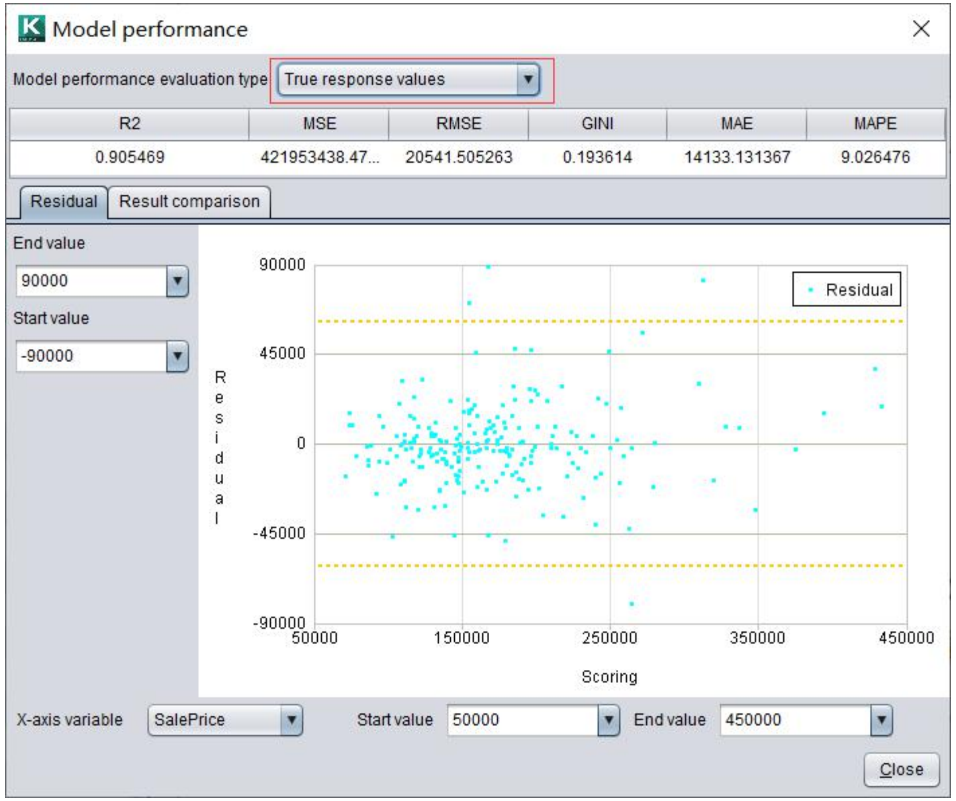


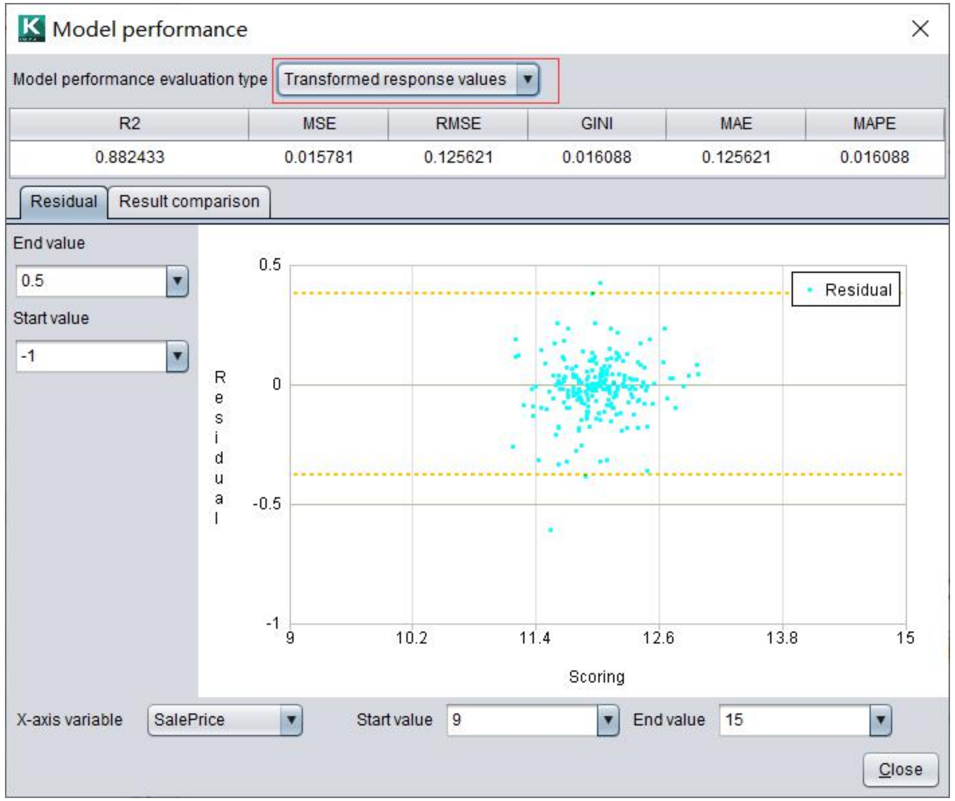


1. Model performance

Regression model: True response values and transformed response values

The performance of regression model can be divided into true value performance and transformed value performance (data value after preprocessed). The true value looks more intuitive, and the transformed value is more accurate for the evaluation of model performance.





1. Model performance

Regression model: evaluation index

Intelligent modeling provides six commonly used evaluation indexes of regression model:

Model performance evaluation type: True response values

R2	MSE	RMSE	GINI	MAE	MAPE
0.905469	421953438.47...	20541.505263	0.193614	14133.131367	9.026476

Model performance evaluation type: Transformed response values

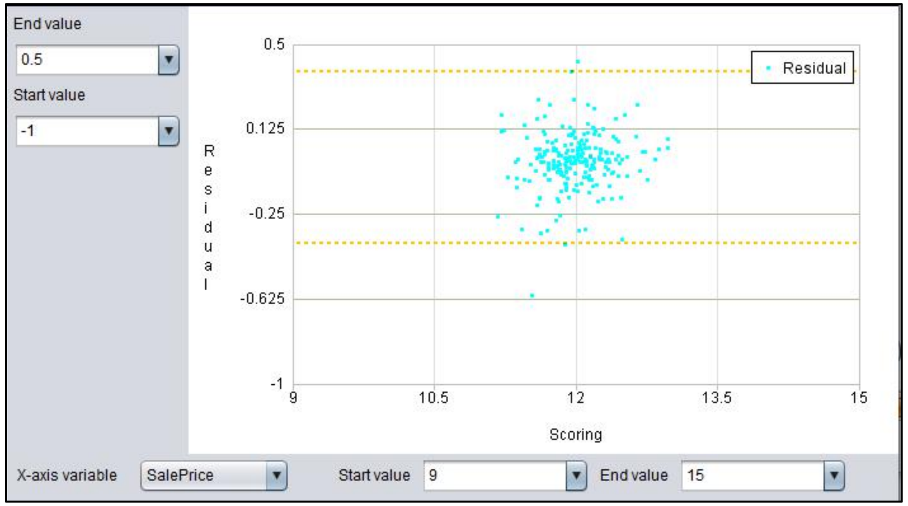
R2	MSE	RMSE	GINI	MAE	MAPE
0.882433	0.015781	0.125621	0.016088	0.125621	0.016088

Evaluation Index	Description
R^2	R^2 is the ratio of the sum of the square of the error between the predicted value and the observed value to the sum of the square of the difference between the observed value and the observed mean value.
MSE	The average sum of the squares of the deviations of the predicted value from the true value.
RMSE	The square root of MSE. The order of magnitude is the same as the true value.
GINI	The average of the absolute value of the deviation between the predicted value and the true value.
MAE	The average of the absolute value of the deviation between the predicted value and the true value.
MAPE	The average of the absolute value of the deviation between the predicted value and the true value.

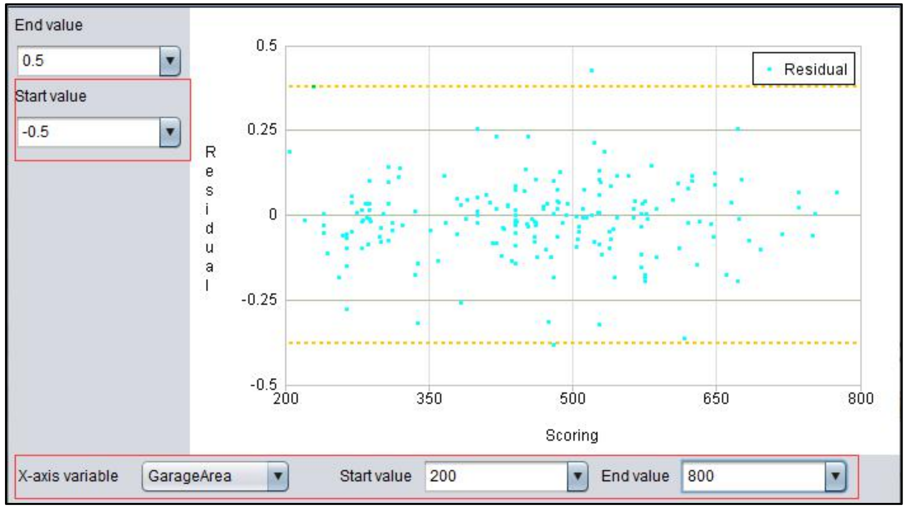
1. Model performance

Regression model: residual chart

The residual is the difference between the observed value and the predicted value. The residual chart is a scatter chart with the residual as the vertical axis and any numerical variable as the horizontal axis. The yellow line is three times RMSE.



You can adjust the horizontal axis variable and the value range of the horizontal and vertical axis for further viewing.

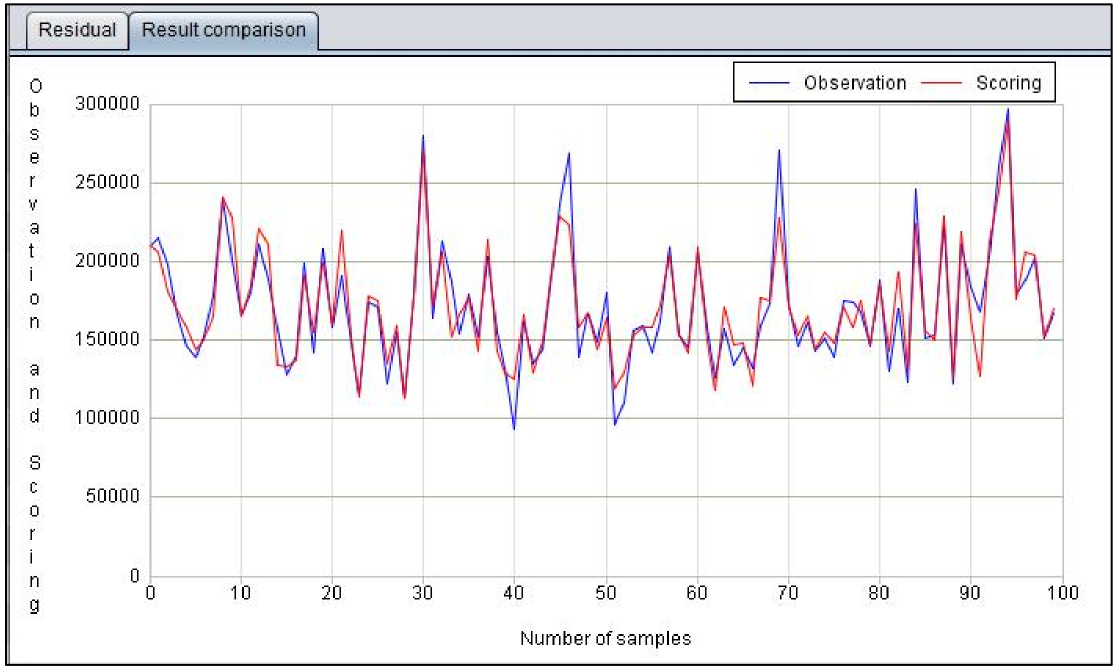


1. Model performance

Regression model: result comparison chart

The horizontal axis of the result comparison chart are the samples of random distribution, and the vertical axis is the corresponding observation value and prediction value.

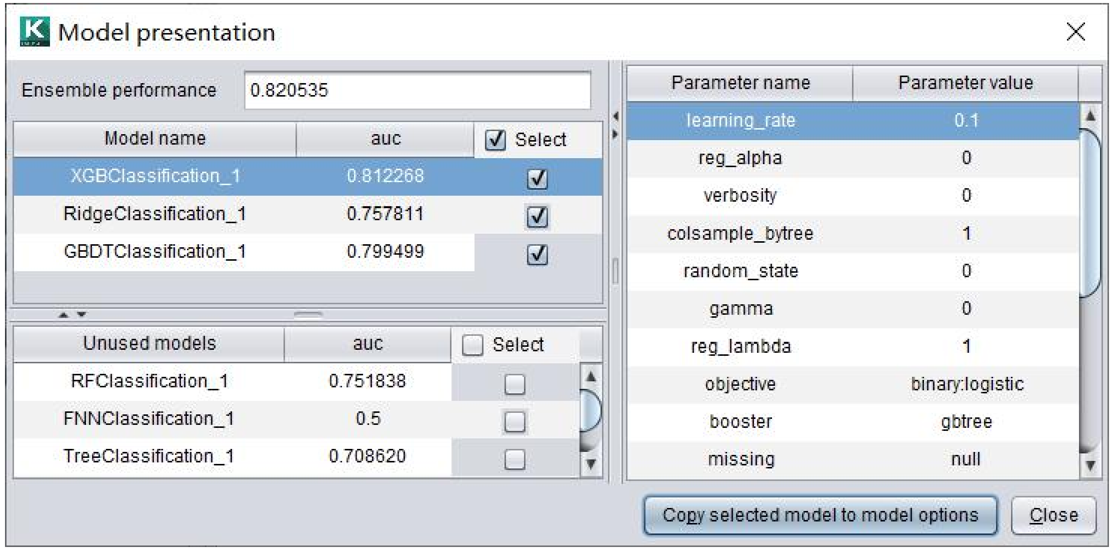
Blue is the observed value and red is the predicted value.



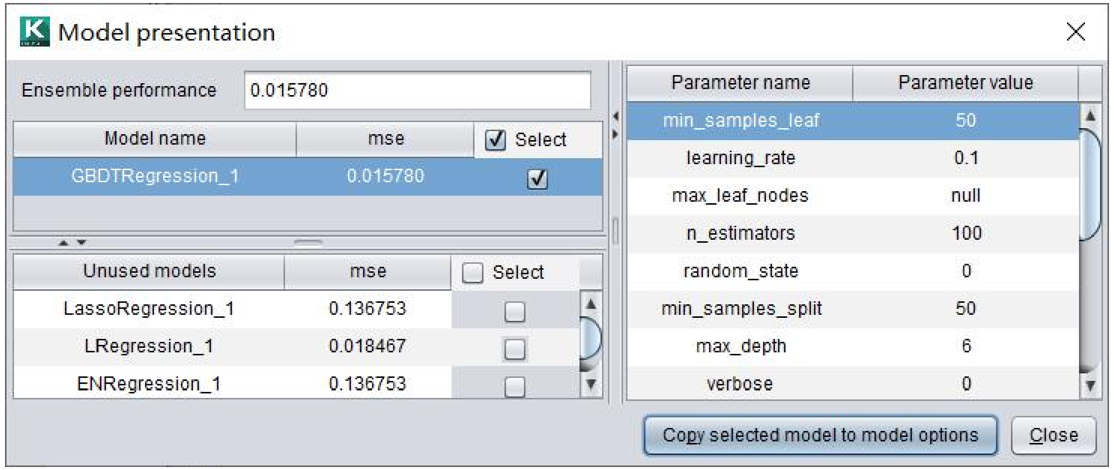
2. Model presentation

The model presentation lists the final selected model combinations and the parameter values of each model.

The selected model parameters can be copied to the model options through the button to further optimize the model parameters.



The final classification model and parameters of Titanic model



The final regression model and parameters of house price model

3. Variable importance

After modeling, the importance information of each variable can be obtained. From the returned importance of Titanic model, we can see that age (children first) and ticket price (higher class) are the most important factors for survival.

Target variable: Survived — Set — Variable filter

NO.	Variable name	Type	Date format	Select	Importance
1	Age	Numerical variable		☑	1
2	Fare	Numerical variable		☑	0.915
3	SibSp	Count variable		☑	0.357
4	PassengerId	ID		☑	0
5	Survived	Binary variable		☑	-
6	Pclass	Categorical variable		☑	0
7	Name	Categorical variable		☐	0
8	Sex	Binary variable		☑	0
9	Parch	Count variable		☑	0
10	Ticket	Categorical variable		☑	0
11	Cabin	Categorical variable		☑	0
12	Embarked	Categorical variable		☑	0

	The role of variable importance
1	Refer to the importance of variables and reprocess the data accordingly.
2	The important variables are used interactively to generate the derived variables, such as distance / time = speed, speed * time = distance and so on.
3	Refer to the importance of variables and make targeted suggestions to customers.

CONTENTS

1. Batch prediction
2. Single prediction

06

Prediction

1. Batch prediction

After you create the model, you can use test data for prediction.

For the binary classification model, the first column is the probability that the target variable is a positive sample.

Taking Titanic as an example, the probability of survival of No. 624 passenger is predicted to be 32.984%.

Batch scoring | Scoring

Scoring data: C:\Program Files\yimming\yimming\data\titanic_test.csv

Survived_1_percentage	PassengerId	Survived	Pclass	Name	Sex
32.984%	624	0	3	Hansen, Mr. Henry Damsgaard	male
33.937%	625	0	3	"Bowen, Mr. David John ""Dai"""	male
34.68%	626	0	1	Sutton, Mr. Frederick	male
30.683%	627	0	2	Kirkland, Rev. Charles Leonard	male
58.263%	628	1	1	Longley, Miss. Gretchen Fiske	female
11.971%	629	0	3	Bostandyeff, Mr. Guentcho	male
5.488%	630	0	3	O'Connell, Mr. Patrick D	male
29.972%	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male
2.183%	632	0	3	Lundahl, Mr. Johan Svensson	male
75.61%	633	1	1	Stahelin-Maeglin, Dr. Max	male
2.658%	634	0	1	Parr, Mr. William Henry Marsh	male
27.029%	635	0	3	Skoog, Miss. Mabel	female
37.865%	636	1	2	Davis, Miss. Mary	female
43.924%	637	0	3	Leinonen, Mr. Antti Gustaf	male
65.287%	638	0	2	Collyer, Mr. Harvey	male
46.579%	639	0	3	Panula, Mrs. Juha (Maria Emilia Ojala)	female
21.965%	640	0	3	Thorneycroft, Mr. Percival	male
34.018%	641	0	3	Jensen, Mr. Hans Peder	male
77.531%	642	1	1	Sagesser, Mlle. Emma	female
31.062%	643	0	3	Skoog, Miss. Margit Elizabeth	female
57.925%	644	1	3	Foo, Mr. Choong	male

1. Batch prediction

For regression model, the first column is the predicted value of the target variable.

Taking the house price as an example, the price of house 1461 is predicted to be 120644.118.

Batch scoring | Scoring

Scoring data C:\Program Files\yimming\yimming\data\house_prices_test.csv

SalePrice_predictvalue	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape
	1461	20	RH	80	11622	Pave		Reg
	1462	20	RL	81	14267	Pave		IR1
	1463	60	RL	74	13830	Pave		IR1
	1464	60	RL	78	9978	Pave		IR1
	1465	120	RL	43	5005	Pave		IR1
	1466	60	RL	75	10000	Pave		IR1
	1467	20	RL		7980	Pave		IR1
	1468	60	RL	63	8402	Pave		IR1
	1469	20	RL	85	10176	Pave		Reg
	1470	20	RL	70	8400	Pave		Reg
	1471	120	RH	26	5858	Pave		IR1
	1472	160	RM	21	1680	Pave		Reg
	1473	160	RM	21	1680	Pave		Reg
	1474	160	RL	24	2280	Pave		Reg
	1475	120	RL	24	2280	Pave		Reg
	1476	60	RL	102	12858	Pave		IR1
	1477	20	RL	94	12883	Pave		IR1
	1478	20	RL	90	11520	Pave		Reg
	1479	20	RL	79	14122	Pave		IR1
	1480	20	RL	110	14300	Pave		Reg
	1481	60	RL	105	13650	Pave		Reg

1. Batch prediction

When the target variable is a categorical variable, the probability (sum of 1) of each target classification value is displayed after prediction. For example, for the first record, the probability of target value of 2 is the highest, which is 97.402%.

Batch scoring | Scoring

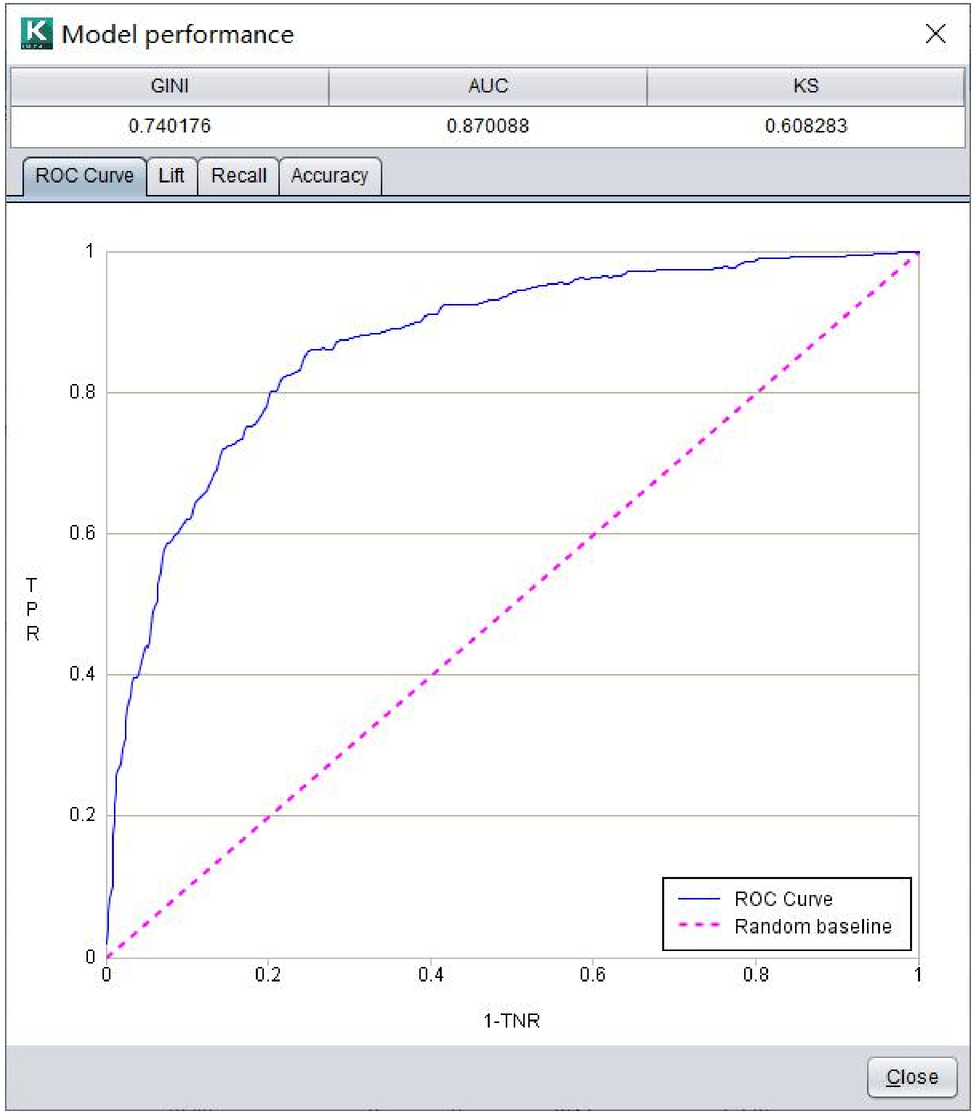
Scoring data: C:\Program Files\yimming\yimming\data\Forest_Covertype.mtx

Cover_Type_1_percentage	Cover_Type_2_percentage	Cover_Type_3_percentage	Cover_Type_4_percentage	Cover_Type_5_percentage	Cover_Type_6_percentage	Cover_Type_7_percentage
0.448%	97.402%	0.169%	0.021%	1.745%	0.177%	0.038%
0.297%	98.152%	0.115%	0.015%	1.223%	0.172%	0.027%
1.875%	97.405%	0.594%	0.01%	0.088%	0.011%	0.017%
3.302%	94.912%	1.172%	0.014%	0.146%	0.429%	0.025%
0.319%	97.864%	0.091%	0.014%	1.546%	0.137%	0.027%
0.768%	96.389%	0.337%	0.034%	2.059%	0.359%	0.054%
0.699%	95.365%	0.171%	0.021%	3.529%	0.176%	0.039%
0.37%	96.957%	0.095%	0.015%	2.385%	0.148%	0.029%
0.511%	97.973%	0.107%	0.014%	1.211%	0.163%	0.021%
0.673%	98.115%	0.073%	0.013%	0.999%	0.103%	0.024%
0.421%	98.58%	0.137%	0.011%	0.708%	0.124%	0.019%
3.994%	95.644%	0.044%	0.022%	0.222%	0.026%	0.047%
2.927%	96.683%	0.178%	0.01%	0.155%	0.028%	0.019%
0.229%	98.33%	0.182%	0.011%	1.119%	0.11%	0.018%
0.318%	98.225%	0.133%	0.024%	0.802%	0.448%	0.05%
0.704%	94.935%	0.224%	0.041%	2.922%	1.099%	0.074%
1.336%	96.347%	0.178%	0.033%	1.74%	0.315%	0.052%
0.383%	96.798%	0.146%	0.027%	2.265%	0.329%	0.053%
0.234%	94.256%	0.108%	0.016%	5.058%	0.297%	0.03%
0.252%	96.695%	0.12%	0.019%	2.536%	0.335%	0.043%
0.681%	97.92%	0.139%	0.029%	0.718%	0.452%	0.06%
6.872%	92.693%	0.026%	0.018%	0.334%	0.021%	0.035%

1. Batch prediction

Generally, the prediction data does not contain the target variable.

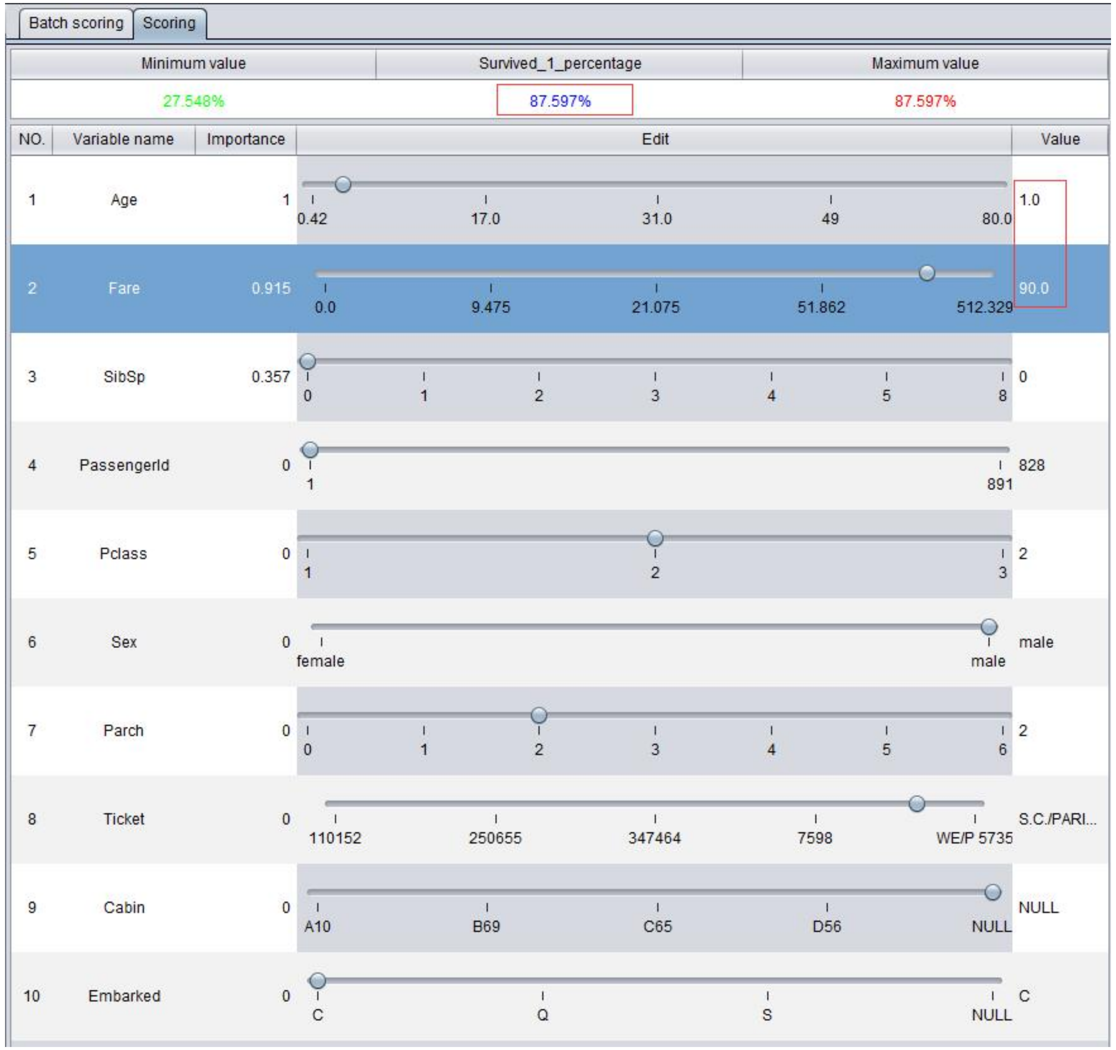
When target variable is included in the prediction data, the performance of the model can be calculated according to the prediction result to evaluate the model.



2. Single prediction

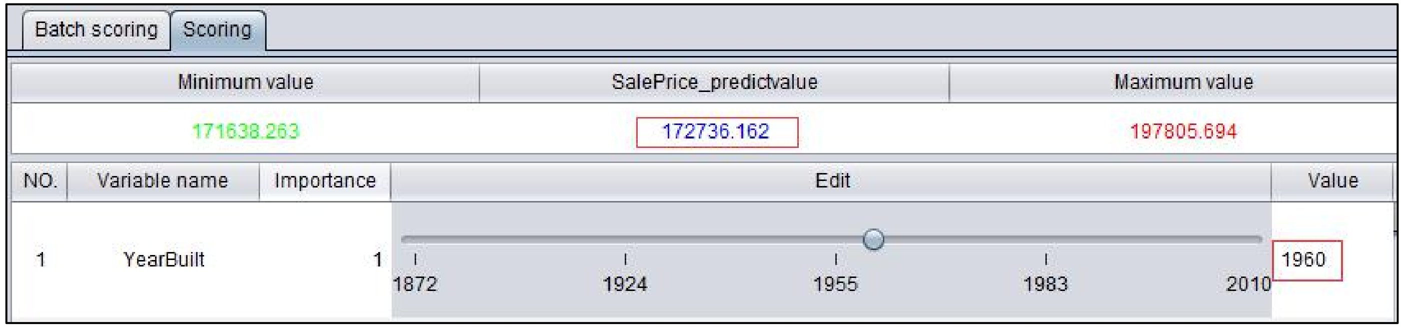
A single prediction can be dragged to modify the variable value and view the prediction result in real time.

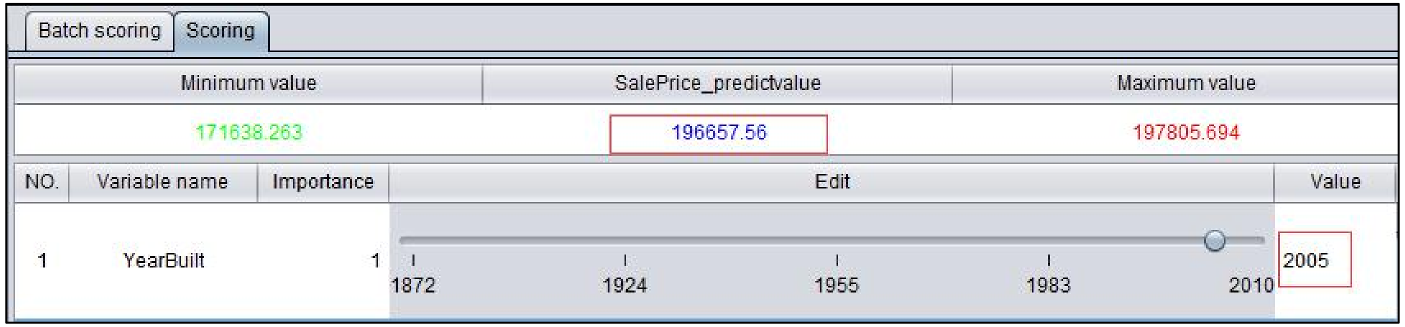
The variables are arranged in descending order of importance, and the top variables usually have more influence on the prediction result. It can be seen that the survival rate of the younger and higher priced tickets is very high.



2. Single prediction

For the house price prediction model, we can see that when the house construction time is dragged from 1960 to 2005 (other variables have not changed), the house price has greatly increased.





CONTENTS

07

Integration solution

1. esProc External library

esProc external library provides interface functions for intelligent modeling, which can be called by SPL. The SPL for modeling:

	A	B
1	=file("titanic_train.csv").cursor@cqt()	/Create training data cursor
2	=ym_env()	/Initialize environment
3	=ym_model(A2,A1)	/Loading data
4	=ym_target(A3, "Survived")	/Set target variable
5	=ym_build_model(A3)	/Execute modeling
6	=ym_save_pcf(A5,"titanic.pcf")	/Save model file
7	=ym_json(A5)	/Export model information as JSON string
8	=ym_importance(A5)	/Get variable importance
9	=ym_present(A5)	/Get model presentation
10	=ym_performance(A5)	/Get model performance
11	>ym_close(A2)	/Close

A7

Value
{"Importance":{"PassengerId":0,"Pclass":0,"Sex":0,""Age":0.433191…

A8

Name	Importance
PassengerId	0.0
Pclass	0.0
…	…

A9

name	value	properties
XGBClass…	0.815	[[max_delt...
XGBClass…	0.777	[[max_delt...
…	…	…

A10

Name	Value
GINI	0.617
AUC	0.808
…	…

For details, please refer to： http://c.raqsoft.com/article/1571711202215

1. esProc External library

After the model is created (or the model created by the intelligent modeling designer), the external library of intelligent modeling can be called through SPL for prediction. The SPL for Prediction:

	A	B
1	=ym_env()	/Initialize environment
2	=ym_load_pcf("titanic.pcf")	/Loading model file
3	=file("titanic_test.csv").import@cqt()	/Loading prediction data
4	=ym_predict(A2,A3)	/Execute prediction, return predicted result object
5	=ym_result(A4)	/Get predicted result sequence table
6	=ym_json(A4)	/When the prediction data is no less than 20 pieces, the model performance JSON information will be exported according to the prediction data evaluation.
7	>ym_close(A1)	/Close

A5

PassengerId	Survived	Pclass	Name	Sex	…
624	0	3	Hansen,…	male	…
625	0	3	Bowen, …	male	…
626	0	1	Sutton, …	male	…
627	0	2	Kirkland…	male	…
…	…	…	…	…	…

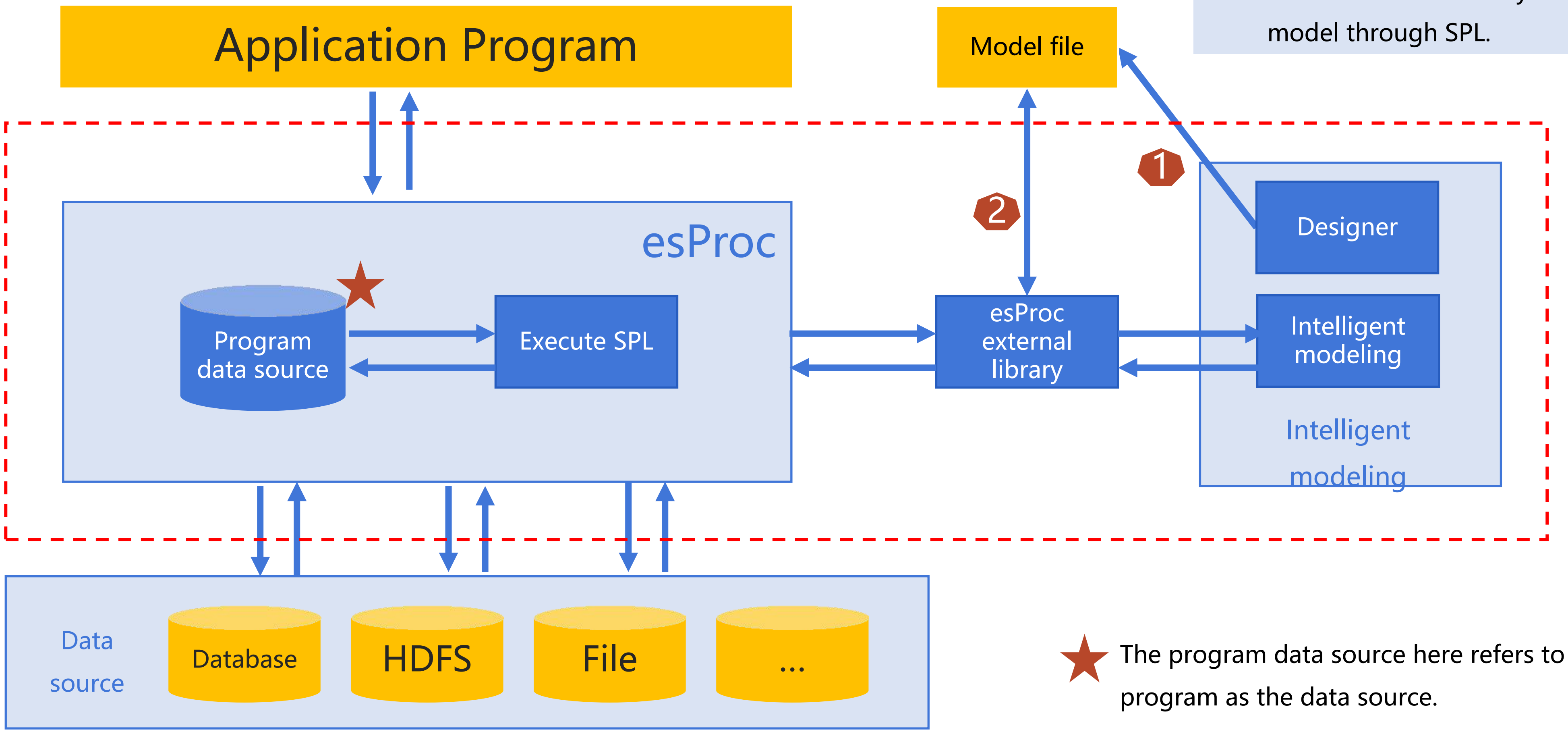
A6

Value
{“Model-Performance”:“{\”GINI\“:0.8369670542635659,\”AUC\“:0.9184835271317829,\”KS\“:0.6867732558139534,\”ROC-Data\“:[\”{\\\“1-specificity\\\”:\\\“0.0\\\”,\\\“sensitivity\\\”:\\\“0.020833333333333332\\\”}\“,\”{\\\“1-…

2. Integration architecture

There are two ways to create a model:

1. Use the intelligent modeling designer to create model file
2. Call the external library of esProc to create model through SPL.



★ The program data source here refers to the SPL program as the data source.

THANKS

Innovation makes progress

www.raqsoft.com.cn