润乾
RAQSOFT

数据挖掘

Titanic幸存者预测

www.raqsoft.com.cn

目录

CONTENTS

目录 CONTENTS

01

数据介绍

数据介绍

Titanic生还者预测是kaggle上的经典赛题，本节以此数据介绍数据挖掘过程
（面向对象是小白，大神请绕过）

“titanic_train.csv” ：
训练集（有目标变量）共有891条记录，12个变量

“titanic_test.csv” ：
待测集（无目标变量）共有418条记录，11个变量

分析目标：
1.寻找影响乘客幸存与否的因素
2.根据训练集构建模型对待测集数据进行预测

序号	变量	描述信息
1	PassengerId	乘客编号
2	Survived	是否幸存
3	Pclass	船票等级
4	Name	乘客姓名
5	Sex	乘客性别
6	Age	乘客年龄
7	SibSp	乘客兄弟姐妹、配偶数量
8	Parch	乘客父母孩、子数量
9	Ticket	船票号码
10	Fare	船票价格
11	Cabin	船舱
12	Embarked	登船港口

数据字典

目录 CONTENTS

02

探索和预处理

前面对数据有了总体了解，下面我们对这个数据集做探索和预处理，尝试找到影响乘客是否幸存的因素。

数据探索

1. 根据数据字典检验变量类型

右图是易明智能建模工具自动识别的变量类型，其中Name被识别为ID，是因为它没有重复值，和PassengerId一样被认为是每条记录的唯一标识了，但我们需要从Name中提取一些信息所以将其改为分类变量。SibSp和Parch表示家庭成员数，但被识别成了分类变量，因此将其改成计数变量。

序号	变量名	类型	日期格式	选出
1	PassengerId	ID		☑
2	Survived	二值变量		☑
3	Pclass	分类变量		☑
4	Name	分类变量		☑
5	Sex	二值变量		☑
6	Age	数值变量		☑
7	SibSp	计数变量		☑
8	Parch	计数变量		☑
9	Ticket	分类变量		☑
10	Fare	数值变量		☑
11	Cabin	分类变量		☑
12	Embarked	分类变量		☑

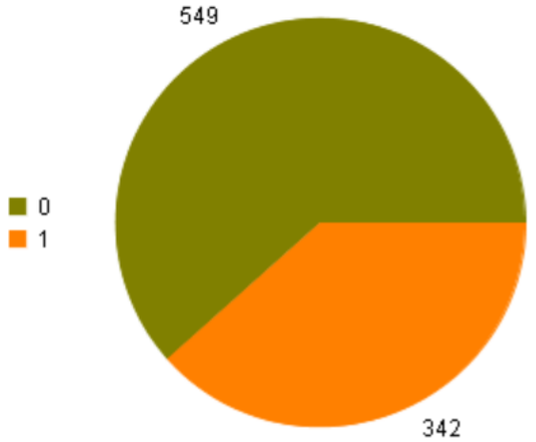
数据探索 - 分布分析

饼图

缺失率	势
0%	2

2. 目标变量Survived

是否幸存，共有两个类别1和0，其中1表示幸存，0表示遇难。没有缺失值，是本次建模的目标变量。



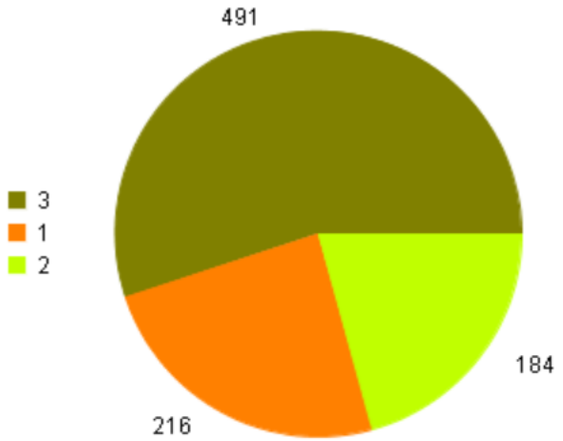
3. Pclass

船票等级，共有三个类别，分别是1、2、3，其中1、2等级人数少，3等级人数多，无缺失值。

按目标变量分组后，查看幸存率，可以发现船票等级越高，幸存率越高。（还是要努力挣钱啊！）

饼图 | 分组目标频数表

缺失率	势
0%	3



饼图 | 分组目标频数表

分类变量	样本量	正样本数	正样本率
1	216	136	62.963%
2	184	87	47.283%
3	491	119	24.236%

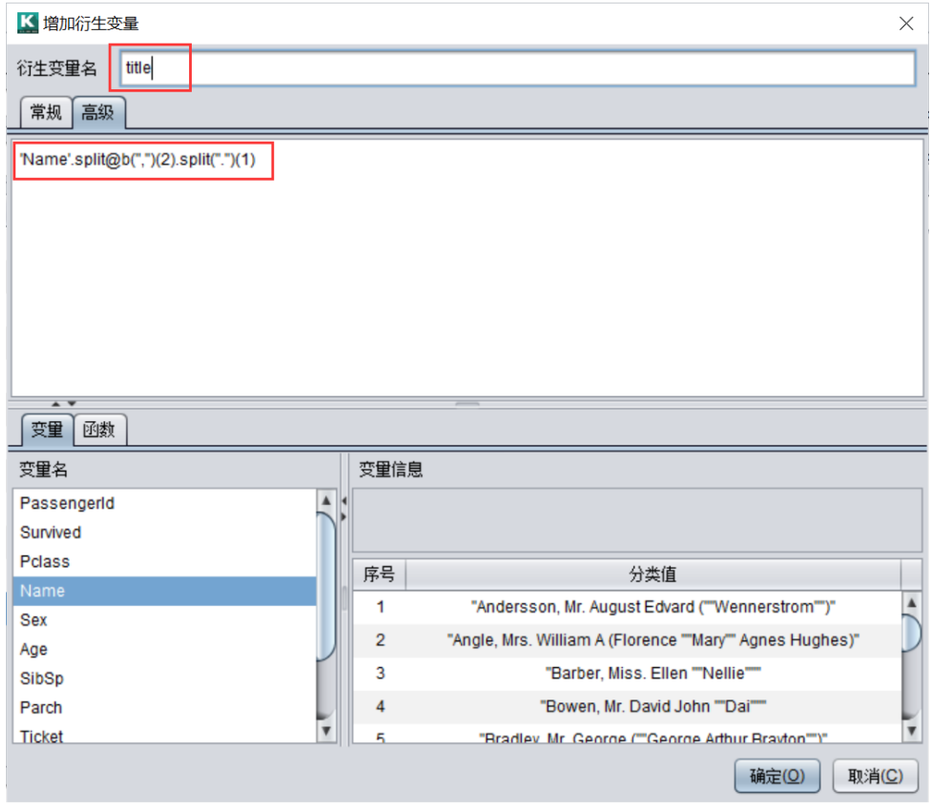
3. Name

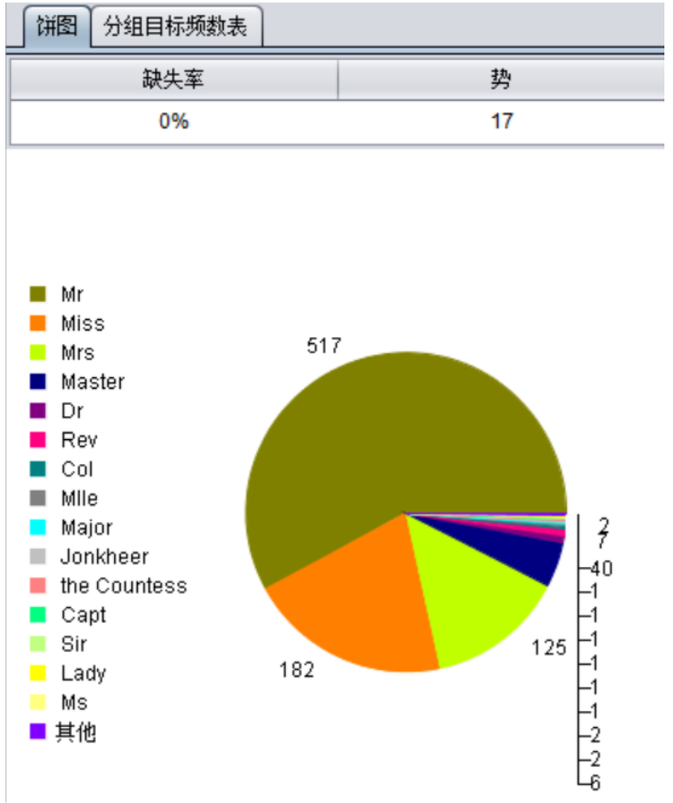
姓名，发现其中包含一些信息，比如人的称呼（mr，miss等），将其提取出来看看有没有用。

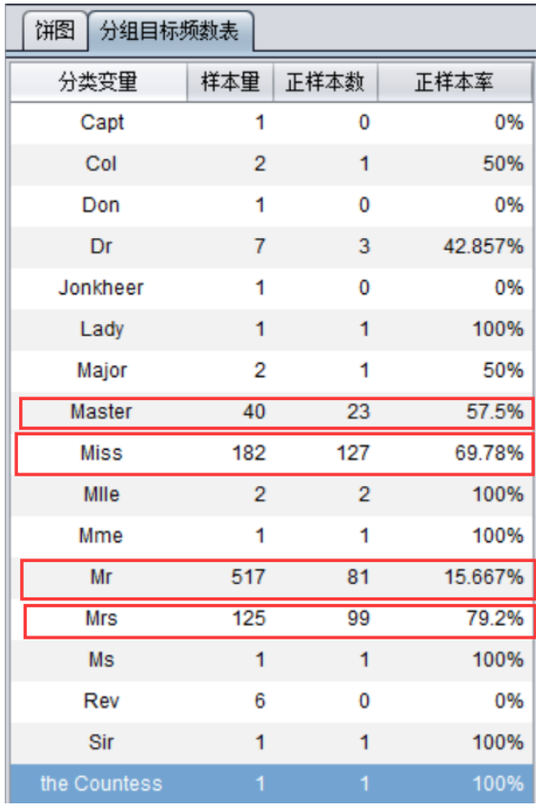
Name
Braund, Mr. Owen Harris
Cumings, Mrs. John Bradley (Florence Briggs Thayer)
Heikkinen, Miss. Laina
Futrelle, Mrs. Jacques Heath (Lily May Peel)
Allen, Mr. William Henry
Moran, Mr. James
McCarthy, Mr. Timothy J
Palsson, Master. Gosta Leonard
Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
Nasser, Mrs. Nicholas (Adele Achem)
Sandstrom, Miss. Marguerite Rut
Bonnell, Miss. Elizabeth
Saundercock, Mr. William Henry
Andersson, Mr. Anders Johan
Vestrom, Miss. Hulda Amanda Adolfina
Hewlett, Mrs. (Mary D Kingcome)
Rice, Master. Eugene
Williams, Mr. Charles Eugene
Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)

数据预处理 - 生成衍生变量

使用变量Name生成变量title







分类变量	样本量	正样本数	正样本率
Capt	1	0	0%
Col	2	1	50%
Don	1	0	0%
Dr	7	3	42.857%
Jonkheer	1	0	0%
Lady	1	1	100%
Major	2	1	50%
Master	40	23	57.5%
Miss	182	127	69.78%
Mlle	2	2	100%
Mme	1	1	100%
Mr	517	81	15.667%
Mrs	125	99	79.2%
Ms	1	1	100%
Rev	6	0	0%
Sir	1	1	100%
the Countess	1	1	100%

从其中提取出姓名的称呼，查看分组后的统计信息，发下Miss，Mrs，Master幸存率很高，但Mr的幸存率很低，说明这个变量是有用的。

数据探索 - 分布分析

4. Sex

性别，共有两个类别男性和女性，其中男性占大多数。无缺失值

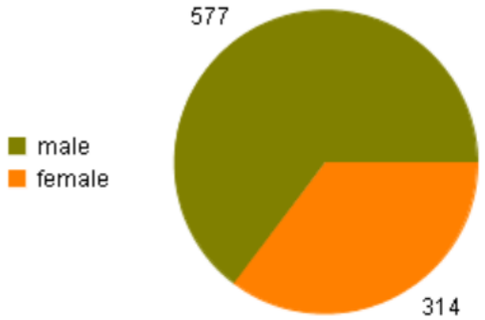
分组后发现女性幸存率远远高于男性。说明这个变量非常重要。

饼图 | 分组目标频数表

缺失率	势
0%	2

饼图 | 分组目标频数表

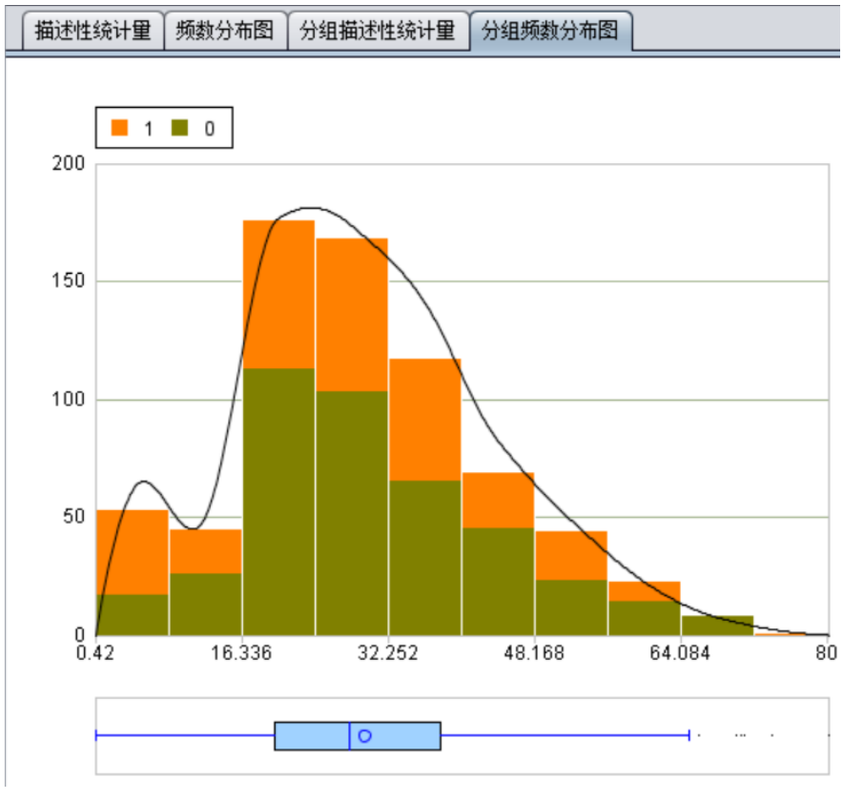
分类变量	样本里	正样本数	正样本率
female	314	233	74.204%
male	577	109	18.891%



5. Age

描述性统计量 | 频数分布图 | 分组描述性统计量 | 分组频数分布图

缺失率	最小值	最大值	平均值	上1/4点	中位数	下1/4点	标准差	偏度
19.865%	0.42	80.0	29.699	38.0	28.0	20.0	14.526	0.388

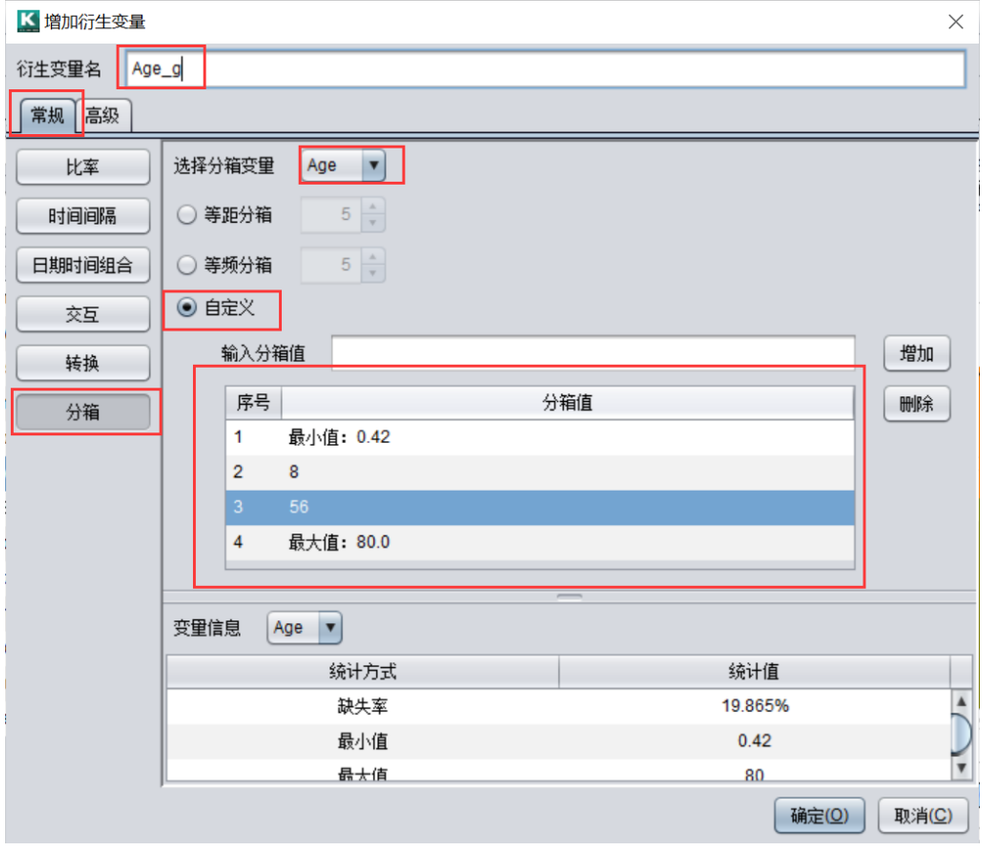


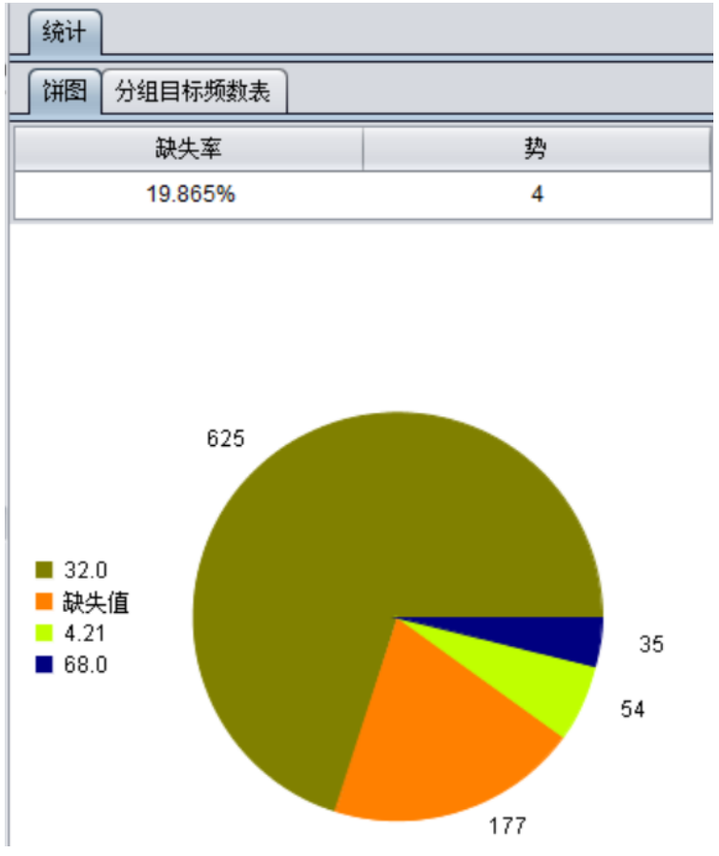
年龄，年龄最小的只有0.42，最大的有80。缺失率是19.865%。智能建模工具会自动进行缺失值填补，不需要处理。

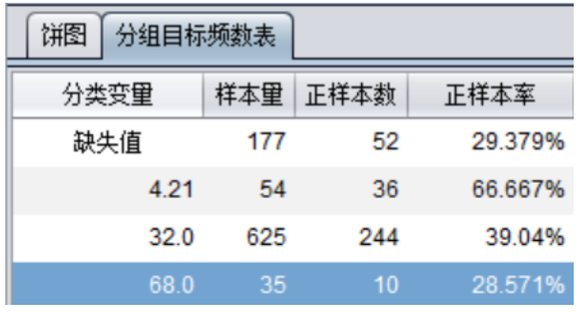
从分组频数分布图看，8岁以下的儿童幸存概率非常大，56岁以上的中老年人幸存概率非常小，青壮年幸存概率变化不大。据此可以将其分为三组，即0~8岁,9~56岁,57岁以上，生成衍生变量Age_g

数据预处理 - 生成衍生变量、连续变量离散化

使用变量Age生成衍生变量Age_g



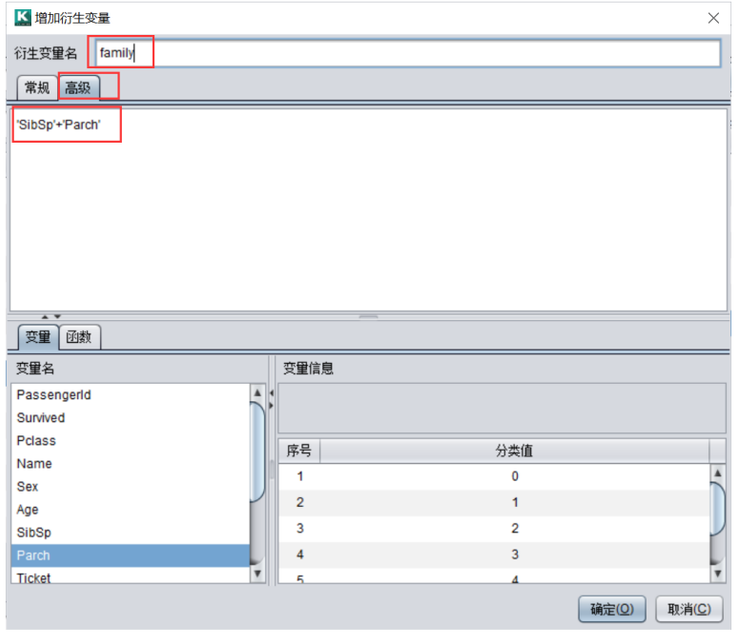


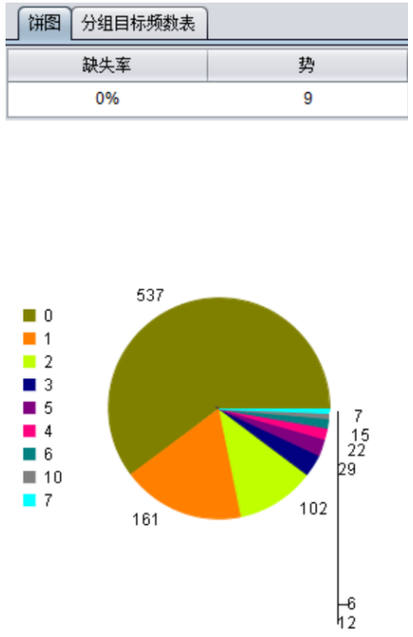


衍生变量Age_g缺失率继承了Age的缺失率，智能建模工具会进行智能补缺，不需要单独处理。查看下分组统计的结果，儿童幸存率高，中老年幸存率低，青壮年幸存率处于两者之间，是个重要变量。

6. 使用SibSp、Parch生成衍生变量family

兄弟姐妹配偶数量、父母子女数量，无缺失值，都是家庭成员，把这两个变量相加组成family变量。





饼图 | 分组目标频数表

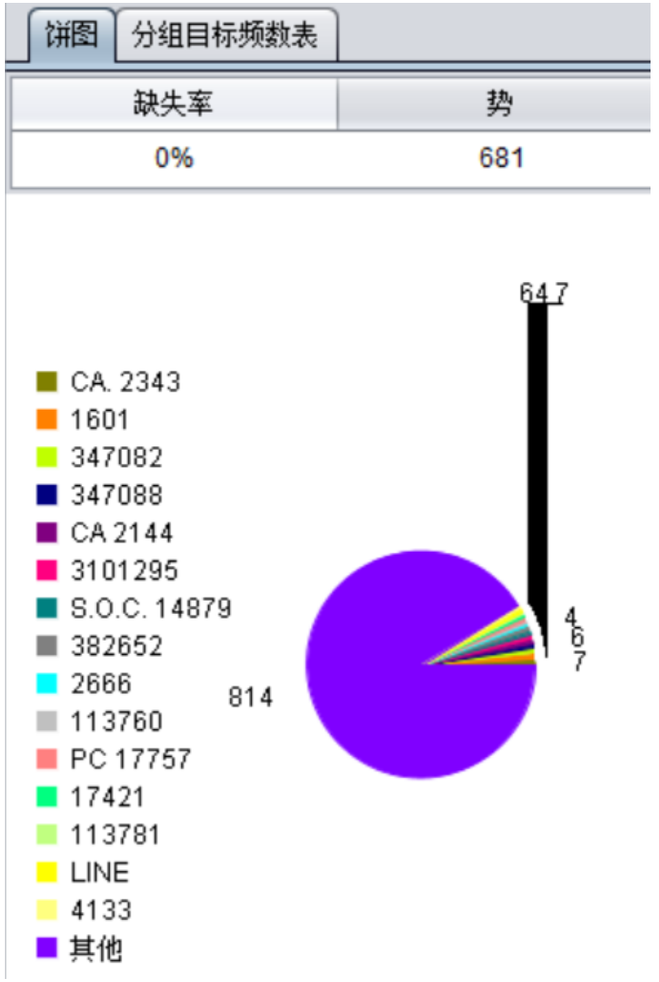
分类变量	样本量	正样本数	正样本率
0	537	163	30.354%
1	161	89	55.28%
2	102	59	57.843%
3	29	21	72.414%
4	15	3	20%
5	22	3	13.636%
6	12	4	33.333%
7	6	0	0%
10	7	0	0%

观察生成的衍生变量family，单身的人占大多数，但幸存率只有30.354%，家庭成员数量在1~3人时，亲情会帮助自己得救，但大于3人时，亲人又会相互牵挂，导致幸存率下降，该变量也是重要变量。

数据探索——分布分析

7. Ticket

船票号，分类数过多，查看饼图和分组统计情况，并不能提供过多的信息，可以将此变量舍弃。



饼图 | 分组目标频数表

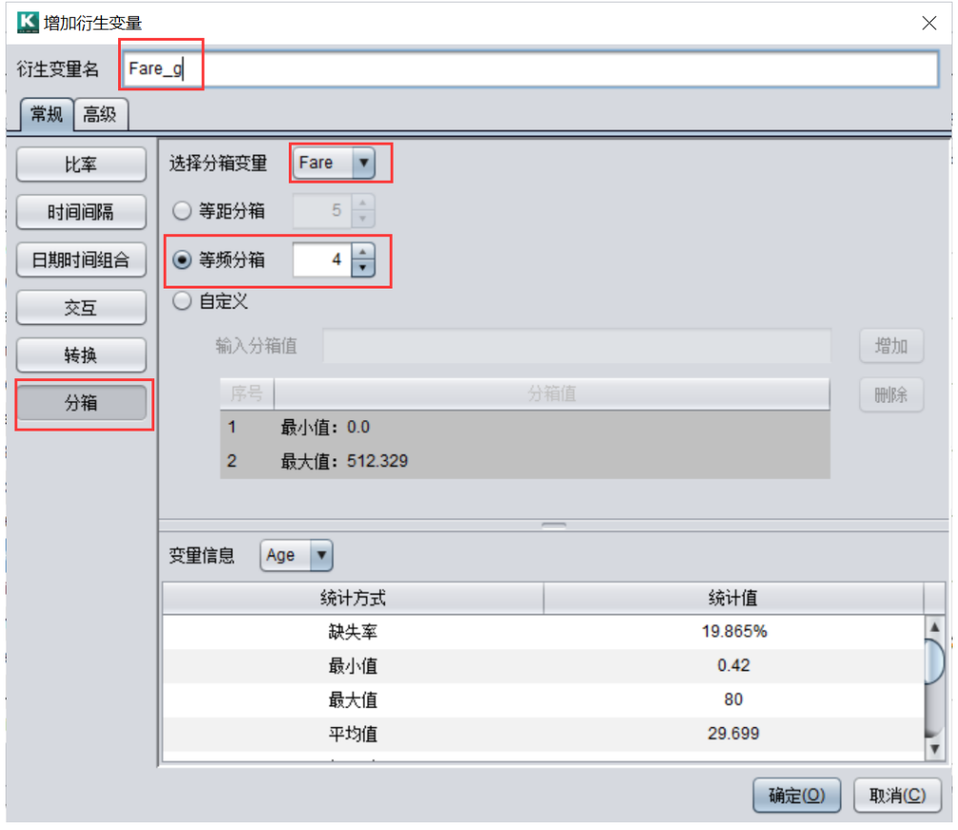
分类变量	样本里	正样本...	正样本率
110152	3	3	100%
110413	3	2	66.667%
110465	2	0	0%
110564	1	1	100%
110813	1	1	100%
111240	1	0	0%
111320	1	0	0%
111361	2	2	100%
111369	1	1	100%
111426	1	1	100%
111427	1	1	100%
111428	1	1	100%
112050	1	0	0%
112052	1	0	0%
112053	1	1	100%
112058	1	0	0%
112059	1	0	0%
112277	1	1	100%

8. Fare

船票价格，最小值0，最大值512.329。偏度4.779，严重右偏，无缺失值。从分布图上可以看出，票价越高，幸存的比例越大，因此可以使用等频分组的方法将其离散化为4个分组。

数据预处理——生成衍生变量、连续变量离散化

将Fare等频离散化为4个分组，生成衍生变量Fare_g

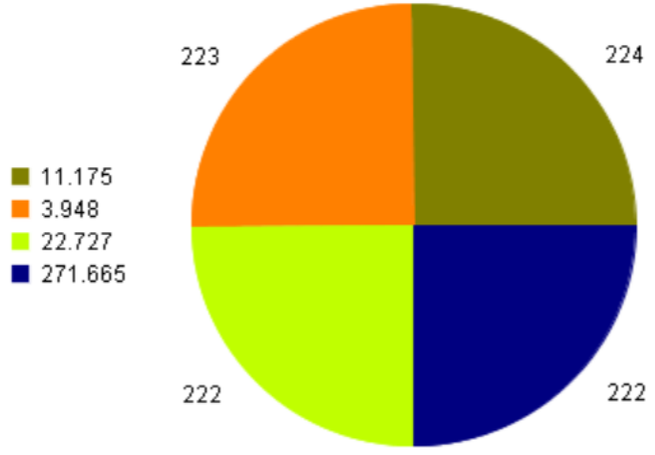


饼图 | 分组目标频数表

缺失率	势
0%	4

饼图 | 分组目标频数表

分类变量	样本量	正样本数	正样本率
3.9479	223	44	19.731%
11.175	224	68	30.357%
22.7271	222	101	45.495%
271.6646	222	129	58.108%



观察生成的衍生变量Fare_g，低票价的分组只有不到20%，而高票价的分组达到了58%。说明此变量可以很好的区分目标变量，是个重要变量。

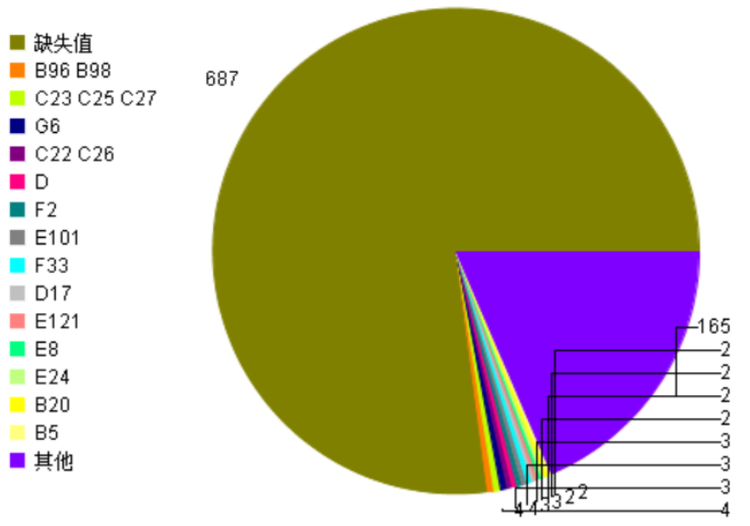
数据探索——缺失值分析、分布分析

9.Cabin

船舱号，分类数很多，缺失率高于77%，看起来这个变量没有用，但我们可以提取该变量是否缺失作为一个信息，即缺失为1，不缺失为0。这也是一种提取缺失值信息的方式。

饼图 | 分组目标频数表

缺失率	势
77.104%	148

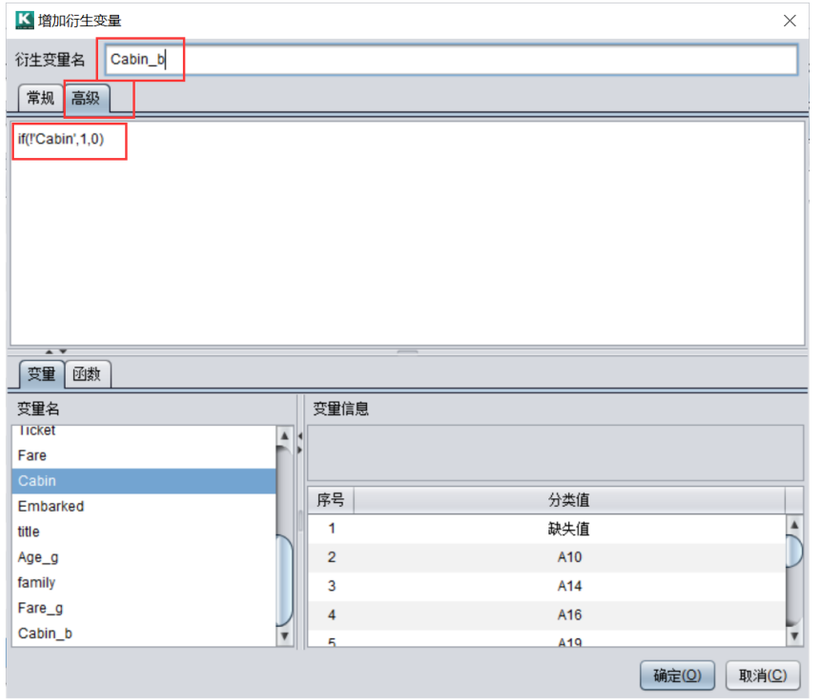


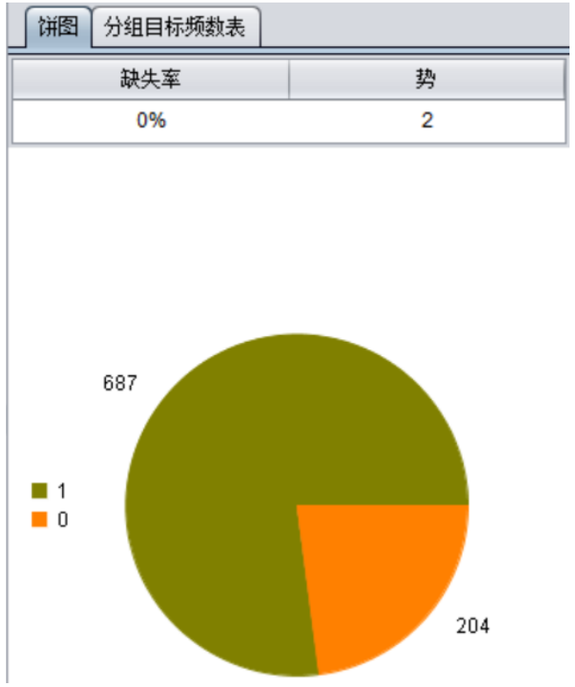
饼图 | 分组目标频数表

分类变量	样本里	正样本数	正样本率
缺失值	687	206	29.985%
A10	1	0	0%
A14	1	0	0%
A16	1	1	100%
A19	1	0	0%
A20	1	1	100%
A23	1	1	100%
A24	1	0	0%
A26	1	1	100%
A31	1	1	100%
A32	1	0	0%
A34	1	1	100%
A36	1	0	0%
A5	1	0	0%

数据预处理——生成衍生变量、提取缺失值信息

提取Cabin的缺失值信息，Cabin缺失为1，不缺失为0，生成衍生变量Cabin_b





分类变量	样本里	正样本数	正样本率
0	204	136	66.667%
1	687	206	29.985%

观察生成的衍生变量Cabin_b，分组的统计信息显示，Cabin不缺失即Cabin=0的幸存率很高，达到了2/3，而缺失的幸存率不足30%。我们可以大胆猜测，好的船舱才有号码，相当于VIP，其他船舱是没有号码的（还是要多挣钱啊）。

数据探索 - 分布分析、对比分析、缺失值分析

10.Embarked

登船港口，共有三个类别，其中S占大多数，C、Q都较少，还有2个缺失值，智能建模工具会自动处理，不需要处理 。

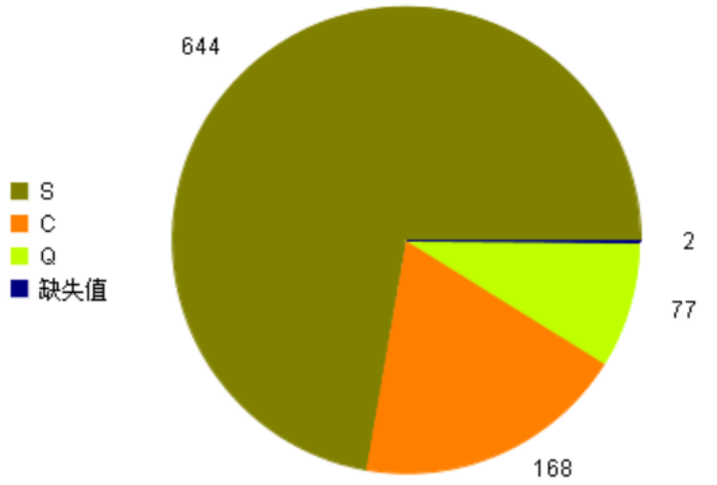
直观上来说登船港口和是否幸存应该没有关系，但数据告诉我们C港口登船的乘客幸存率明显高于其他港口，所以有时直觉这东西并不可靠。

饼图 | 分组目标频数表

缺失率	势
0.224%	4

饼图 | 分组目标频数表

分类变量	样本量	正样本数	正样本率
缺失值	2	2	100%
C	168	93	55.357%
Q	77	30	38.961%
S	644	217	33.696%



数据预处理——变量选择

11.去除无关变量，保留有用变量

1.passengerId，乘客唯一标识，没用，去除；

2.Name，提取了title变量，没用了，去除；

3.Age，生成了Age_g变量，不需要了，去除；

4.SibSp、Parch，生成了family变量，不需要了，去除；

5.Ticket，分类数过多，没用，去除；

6.Fare，生成了Fare_g变量，不需要了，去除；

7.Cabin，生成了Cabin_g变量，不需要了，去除。

序号	变量名	类型	日期格式	选出
1	PassengerId	ID		☐
2	Survived	二值变量		☑
3	Pclass	分类变量		☑
4	Name	分类变量		☐
5	Sex	二值变量		☑
6	Age	数值变量		☐
7	SibSp	分类变量		☐
8	Parch	分类变量		☐
9	Ticket	分类变量		☐
10	Fare	数值变量		☐
11	Cabin	分类变量		☐
12	Embarked	分类变量		☑
13	title	分类变量		☑
14	Age_g	分类变量		☑
15	family	分类变量		☑
16	Fare_g	分类变量		☑

探索和预处理方法汇总

序号	变量名	探索内容	探索结果	预处理内容	预处理结果
1	PassengerId	ID变量，没有有用信息	无用变量，舍弃		
2	Survived	分布分析	正负样本比例接近3:5		
3	Pclass	分布分析，分组统计	等级越高，幸存率越低		
4	Name	内容分析	可以提取出称呼信息	提取有价值信息	生成衍生变量title
5	Sex	分布分析，分组统计	女性幸存率远远高于男性		
6	Age	缺失值分析，分布分析	儿童幸存率高，老人幸存率低	连续变量离散化	生成衍生变量Age_g
7	SibSp	意义分析	兄弟姐妹，配偶，孩子，父母都是家人	变量交互	生成衍生变量family
8	Parch				
9	Ticket	分布分析	没有有用信息，舍弃		
10	Fare	分布分析，分组统计	偏斜严重，票价越高，幸存率越高	连续变量离散化	生成衍生变量Fare_g
11	Cabin	缺失值分析，分布分析	缺失率很高，肯能存在有用信息	提取缺失值信息	生成衍生变量Cabin_b
12	Embarked	缺失值分析，分布分析	C港口登船乘客生存率高		

注意：我们探索时进行了缺失值分析，但没有进行预处理，是因为智能建模工具可以自动智能的进行缺失值处理，因此我们的预处理内容并不包括缺失值预处理。

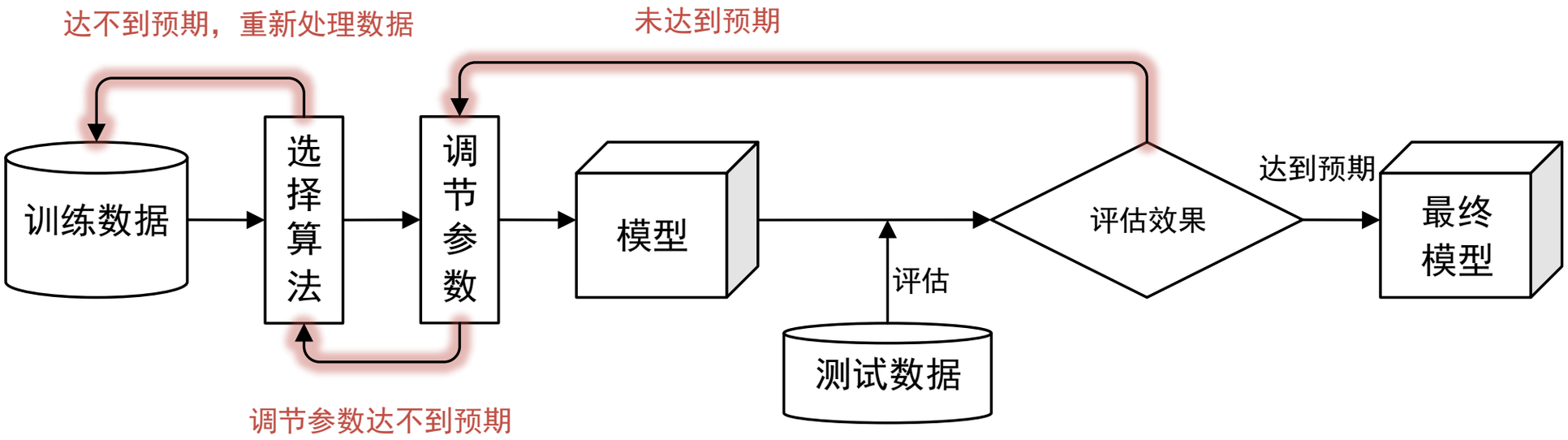
目录 CONTENTS

03

建模与评价

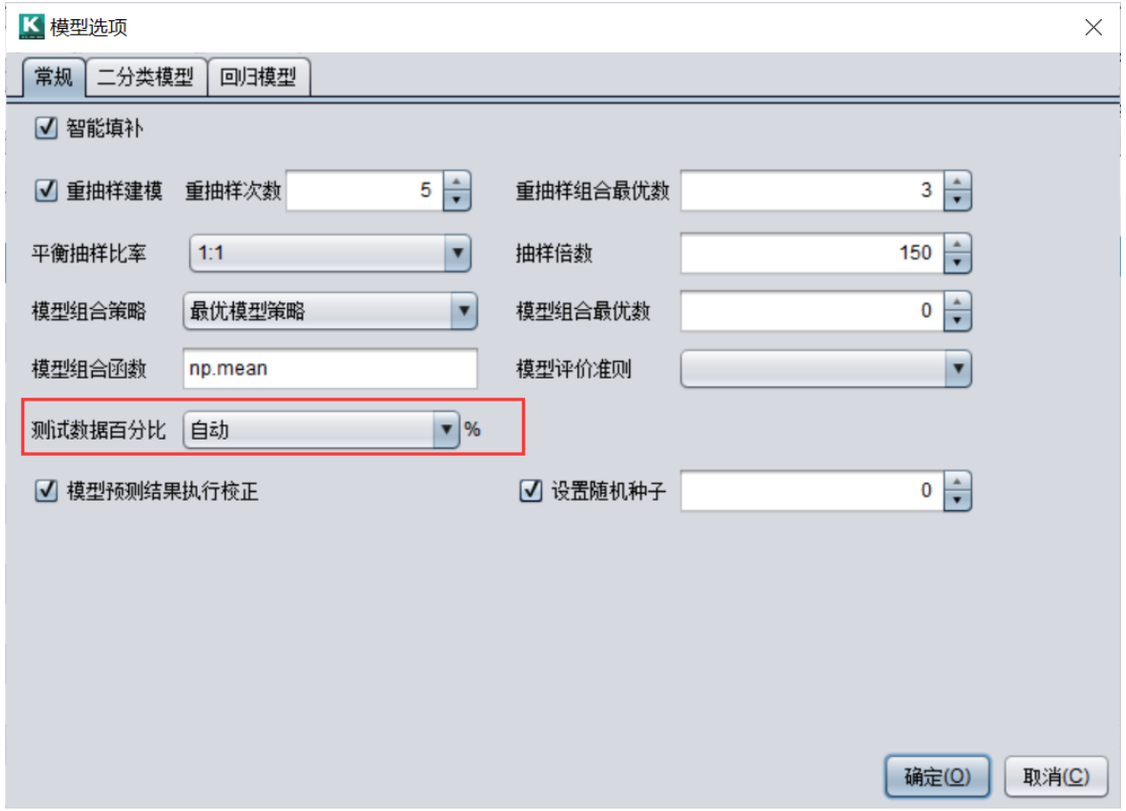
建模与评价

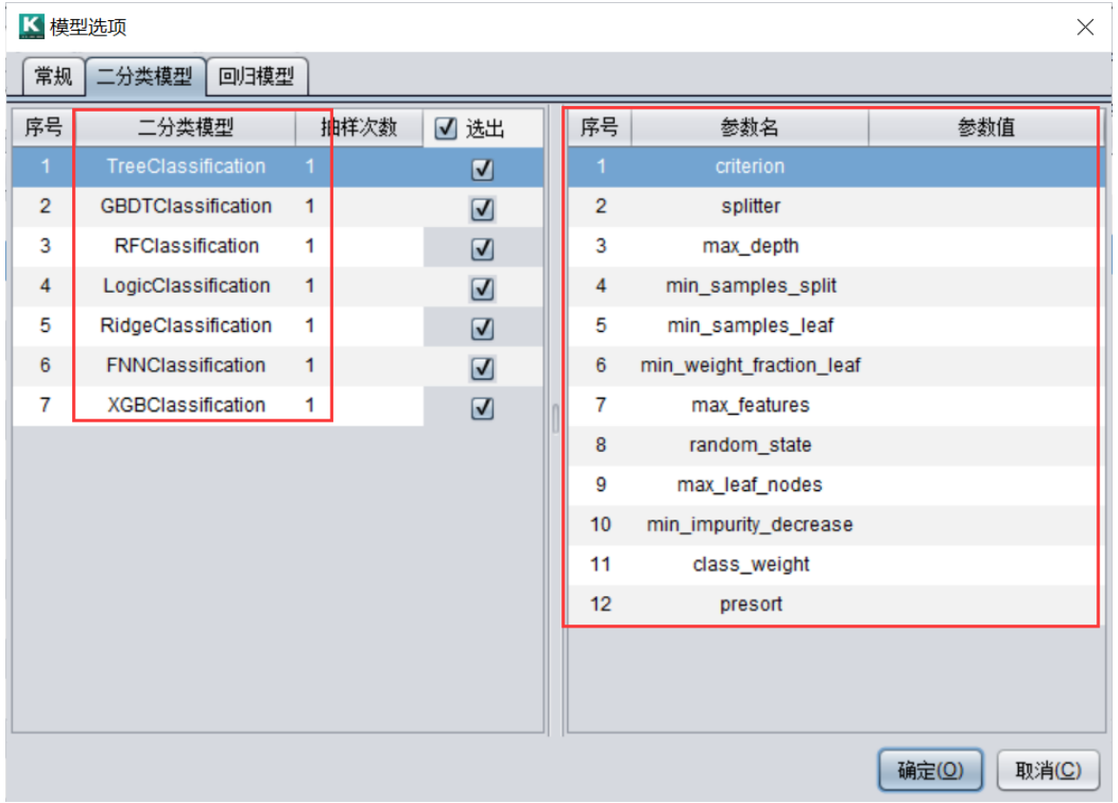
数据预处理后，可以进行建模，建模过程如下图：



建模过程很复杂，我们交给智能建模工具来做，它包括了数据探索，数据预处理，建模、评估等模块。建模时，它会自动将数据切分成训练集和测试集，并在训练集上建模，测试集上评估。它还包括了参数调节，算法融合等智能化建模手段，用户可以很低成本地建成一个比较理想的模型。

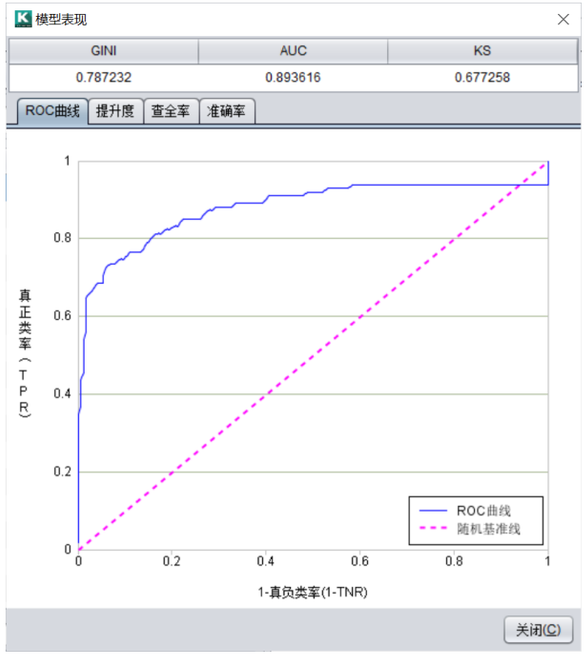
选择模型和调节参数过于复杂，我们只要使用默认的选项就可以了。如下图：



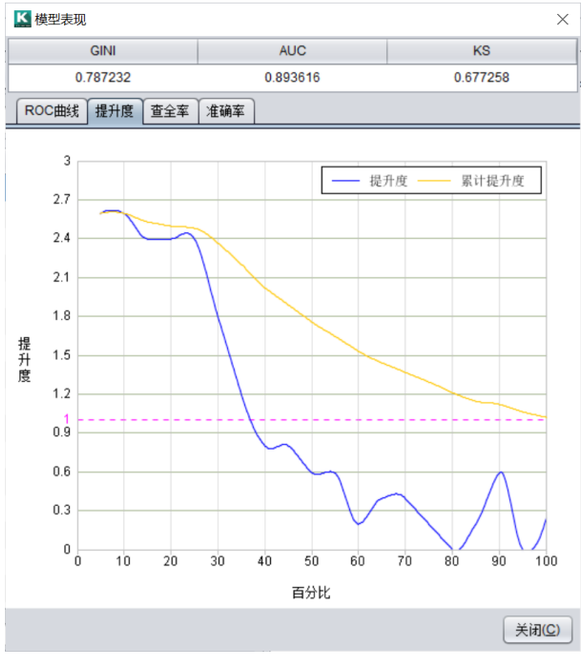


智能建模工具会帮助我们快速的建好模型，采用合适的方法避免过拟合，并在测试集上计算评估指标，方便我们进行评估。

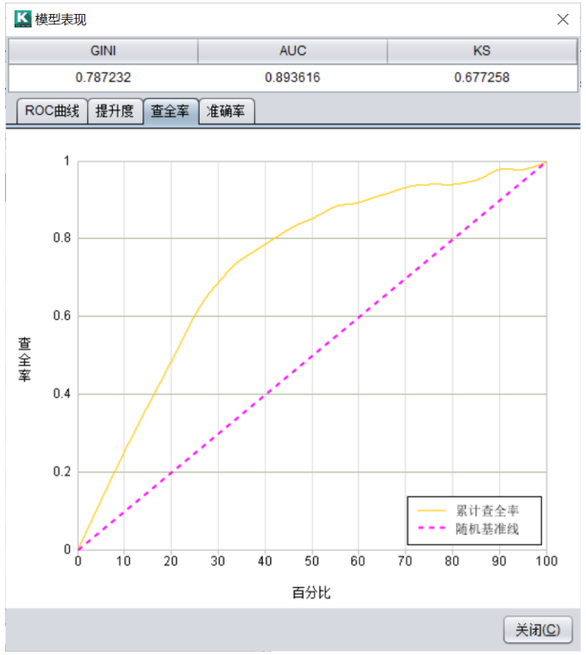
模型表现：



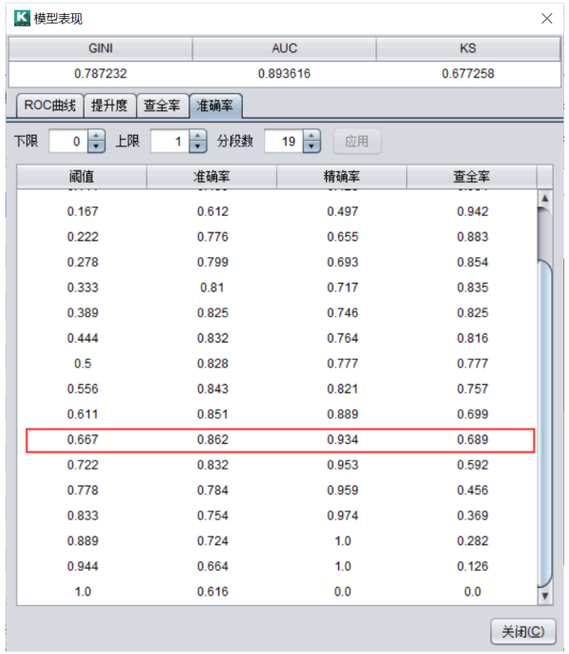
ROC曲线



提升度图



查全率图



准确率表

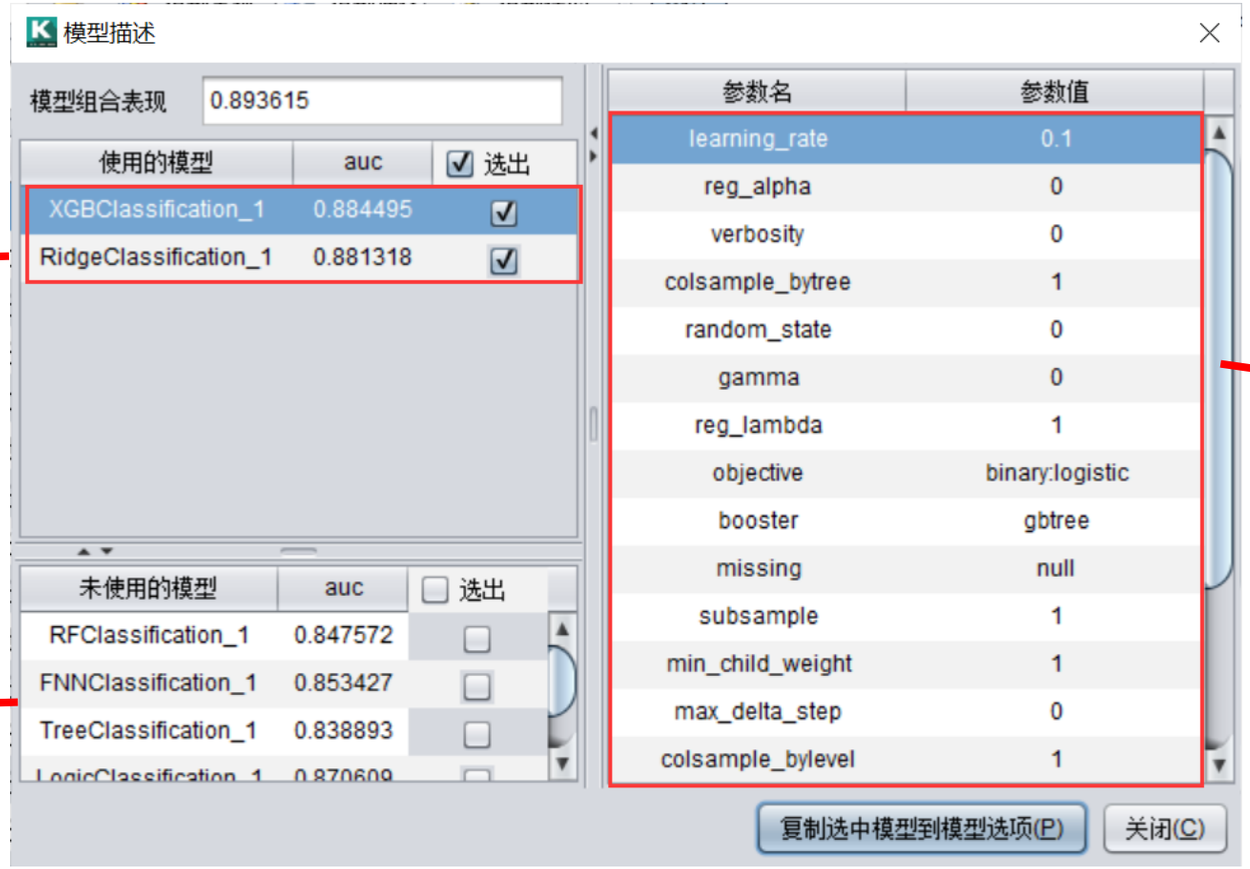
从评价指标来看，模型的表现很不错，其中GINI=0.7872，AUC=0.8936，模型可接受。

观察准确率表可以看到，当阈值取0.667时，准确率最高为0.862。表示把预测概率大于0.667的乘客作为幸存者，小于0.667的乘客作为遇难者时，预测的准确率是最高的。这时的精确率是0.934，表示预测是幸存者的乘客当中，93.4%的乘客确实是幸存者。查全率是0.689，表示预测是幸存者的乘客占全部幸存乘客的68.9%。

建模使用了哪几个模型：

选出的模型

未选出的模型



模型参数

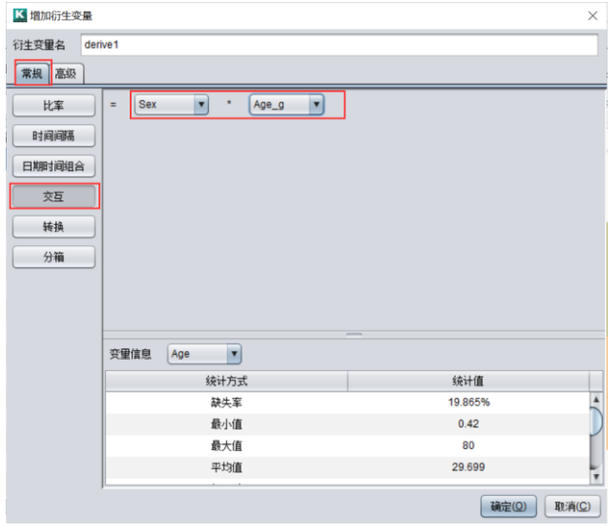
本次建模选出了XGB和Ridge两个分类模型，使用这两个模型组合得到最优的组合模型。模型参数是智能建模自动筛选出的参数，数据挖掘专家可以选择“复制选中模型到模型选项”，修改参数重新建模。

变量重要度：

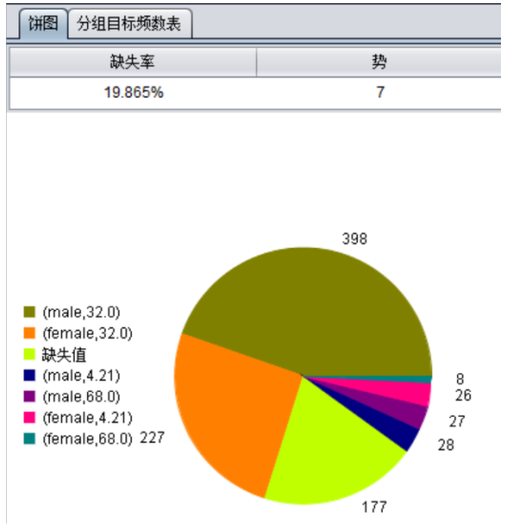
该模型的变量重要度降序排序结果：Sex、Age_g、titile、Pclass、Cabin_b、Fare_g、Embarked。其中衍生变量的重要度都挺高的，说明我们的数据探索和预处理是有效的。

序号	变量名	类型	日期格式	选出	重要度
1	Sex	二值变量		☑	1
2	Age_g	分类变量		☑	0.787
3	family	分类变量		☑	0.604
4	title	分类变量		☑	0.42
5	Pclass	分类变量		☑	0.348
6	Cabin_b	二值变量		☑	0.216
7	Fare_g	分类变量		☑	0.188
8	Embarked	分类变量		☑	0.154
9	PassengerId	ID		☐	0
10	Survived	二值变量		☑	-
11	Name	分类变量		☐	0
12	Age	数值变量		☐	0
13	SibSp	分类变量		☐	0
14	Parch	分类变量		☐	0
15	Ticket	分类变量		☐	0
16	Fare	数值变量		☐	0

使用高重要度变量交互生成衍生变量，如derive1=Sex*Age_g



序号	变量名	类型	日期格式	选出
8	Embarked	分类变量		☑
9	PassengerId	ID		☐
10	Survived	二值变里		☑
11	Name	分类变里		☐
12	Age	数值变里		☐
13	SibSp	分类变里		☐
14	Parch	分类变里		☐
15	Ticket	分类变里		☐
16	Fare	数值变里		☐
17	Cabin	分类变里		☐
18	derive1	分类变里		☑

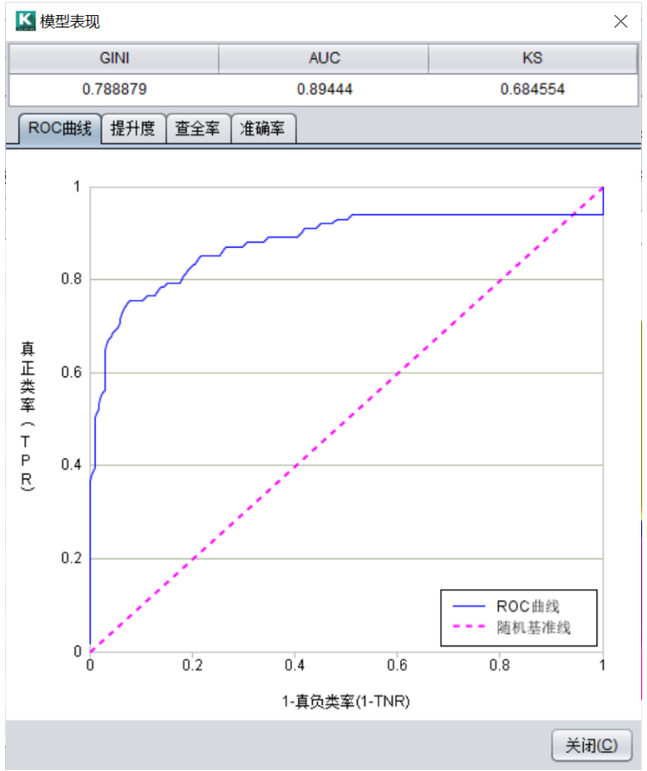


饼图 | 分组目标频数表

分类变量	样本量	正样本数	正样本率
缺失值	177	52	29.379%
(female,32.0)	227	171	75.33%
(female,4.21)	26	19	73.077%
(female,68.0)	8	7	87.5%
(male,32.0)	398	73	18.342%
(male,4.21)	28	17	60.714%
(male,68.0)	27	3	11.111%

还可以增加其他的交互变量，但要注意，当某一个变量类别数很多时（比如family有9个类别）就不适合继续交互了，因为交互后的类别数是两个变量类别数的乘积（如Sex*family的类别数就是18），类别数过多会影响模型效果。建议把family进一步分箱（如按成员数量分成，0,1~3,3个以上三类），然后再进行交互。

增加衍生变量后的模型表现：



目标变量 Survived 设置 筛选变量

序号	变量名	类型	日期格式	选出	重要度
1	derive1	分类变量		☑	1
2	Sex	二值变量		☑	0.374
3	Age_g	分类变量		☑	0.278
4	family	分类变量		☑	0.241
5	Pclass	分类变量		☑	0.185
6	title	分类变量		☑	0.133
7	Fare_g	分类变量		☑	0.121
8	Cabin_b	二值变量		☑	0.098
9	Embarked	分类变量		☑	0.094
10	PassengerId	ID		☐	0
11	Survived	二值变量		☑	-
12	Name	分类变量		☐	0
13	Age	数值变量		☐	0
14	SibSp	分类变量		☐	0
15	Parch	分类变量		☐	0
16	Ticket	分类变量		☐	0

	AUC	GINI
增加衍生变量前	0.893616	0.787232
增加衍生变量后	0.89444	0.788879

增加变量后，模型表现比原来更好了，而且看变量重要度，derive1也成为最重要的变量，说明新增加的衍生变量是有用的。

目录 CONTENTS

02

模型应用

模型应用

模型建好后，使用模型对待测集进行预测。

智能建模工具会自动处理待测集，生成衍生变量，如下图：

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292		Q
893	3	Wilkes, Mrs. James (El...	female	47.0	1	0	363272	7.0		S
894	2	Myles, Mr. Thomas Fra...	male	62.0	0	0	240276	9.6875		Q
895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625		S
896	3	Hirvonen, Mrs. Alexand...	female	22.0	1	1	3101298	12.2875		S

Survived_1_percentage	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	title	Age_g	family	Fare_g	Cabin_b
	892	3	Kelly, Mr...	male	34.5	0	0	330911	7.8292		Q	Mr	32.0	0	3.9479	1
	893	3	Wilkes, ...	female	47.0	1	0	363272	7.0		S	Mrs	32.0	1	3.9479	1
	894	2	Myles, M...	male	62.0	0	0	240276	9.6875		Q	Mr	68.0	0	11.175	1
	895	3	Wirz, Mr....	male	27.0	0	0	315154	8.6625		S	Mr	32.0	0	11.175	1
	896	3	Hirvone...	female	22.0	1	1	3101298	12.2875		S	Mrs	32.0	2	11.175	1

预测结果列

增加的衍生变量

模型应用

模型预测结果见下图：

Survived_1_percentage	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	title	Age_g	family	Fare_g	Cabin_b
11.29%	892	3	Kelly, Mr. ...	male	34.5	0	0	330911	7.8292		Q	Mr	32.0	0	3.9479	1
72.935%	893	3	Wilkes, M...	female	47.0	1	0	363272	7.0		S	Mrs	32.0	1	3.9479	1
9.036%	894	2	Myles, Mr....	male	62.0	0	0	240276	9.6875		Q	Mr	68.0	0	11.175	1
21.742%	895	3	Wirz, Mr. ...	male	27.0	0	0	315154	8.6625		S	Mr	32.0	0	11.175	1
71.484%	896	3	Hirvonen,...	female	22.0	1	1	3101298	12.2875		S	Mrs	32.0	2	11.175	1

预测结果，幸存（Survived=1）的概率，如第一条预测结果是11.29%，表示该乘客只有11.29%的可能性幸存。

润乾
RAQSOFT

THANKS

创新技术 推动应用进步

www.raqsoft.com.cn