数据挖掘

基本概念与过程

www.raqsoft.com.cn

目录

CONTENTS

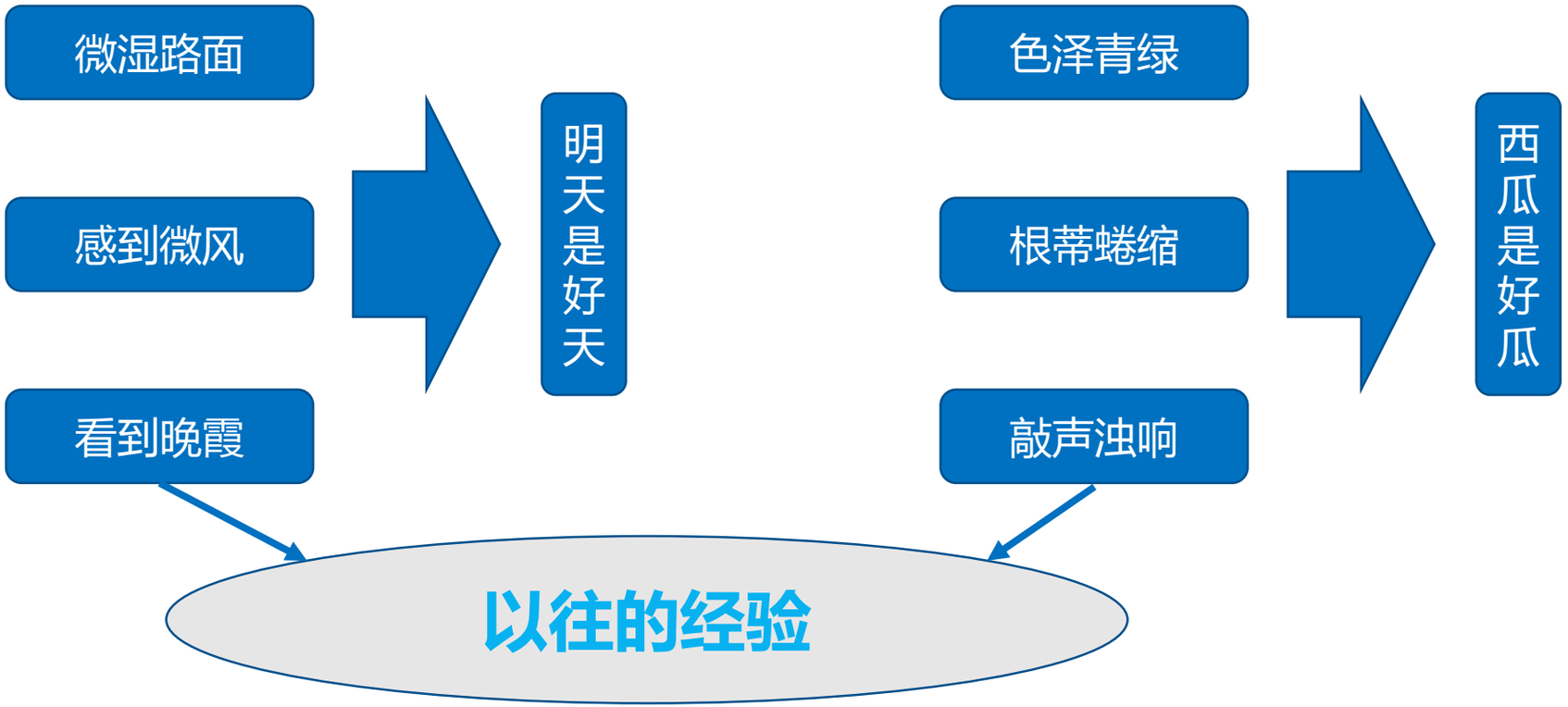
目录 CONTENTS

01

数据挖掘理解

数据挖掘理解

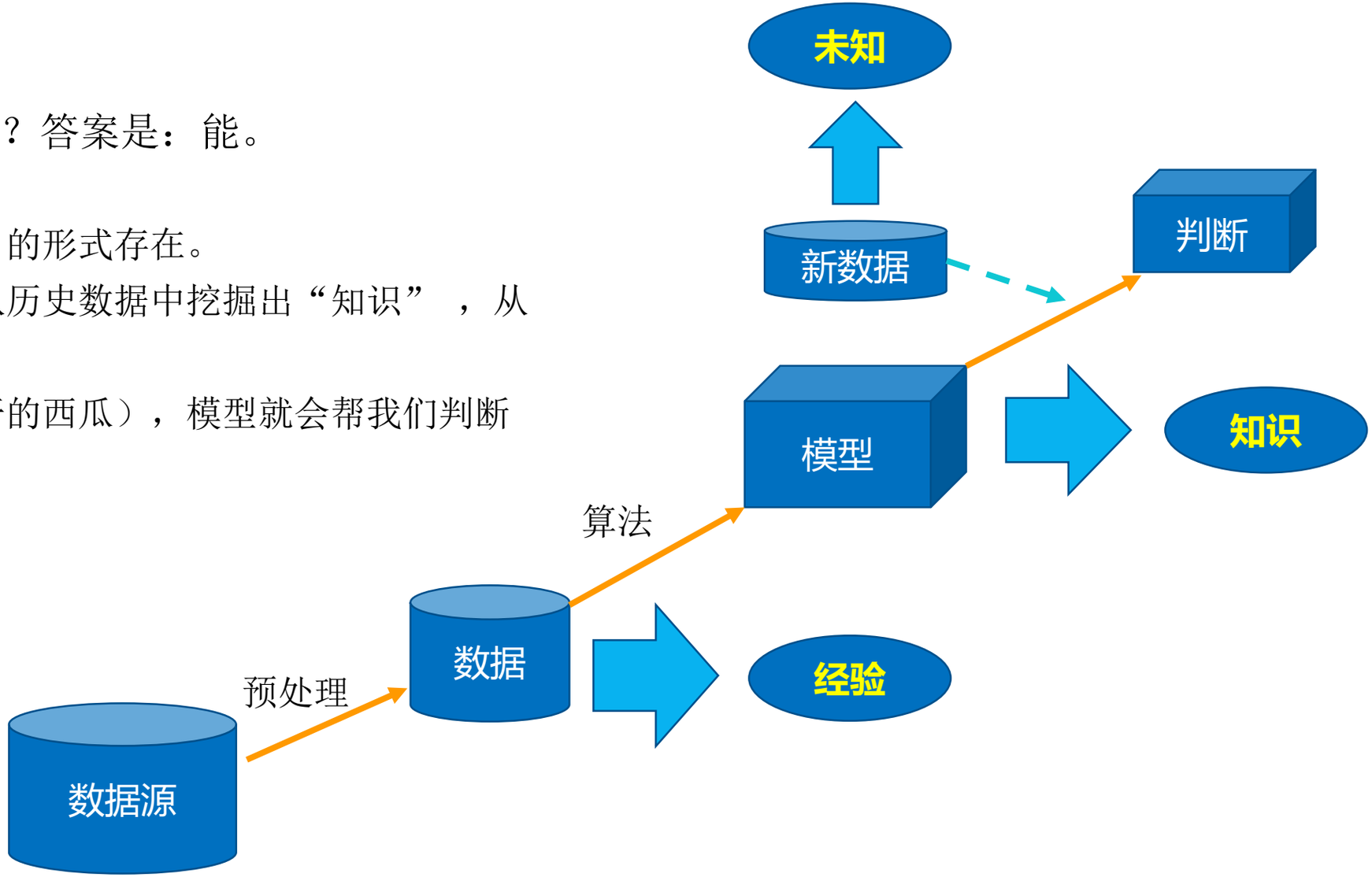
傍晚小街路面上沁出微雨后的湿润，和煦的细风吹来，抬头看看天边的晚霞，嗯，明天又是一个好天气。走到水果摊旁，挑了个根蒂蜷缩、敲起来声音浊响的青绿西瓜，心里期待着享受这个好瓜。



数据挖掘理解

机器能帮我们完成这些吗？答案是：能。

- “经验”通常以“数据”的形式存在。
- 数据挖掘的主要内容是从历史数据中挖掘出“知识”，从而创建“模型”。
- 在面对新情况时（未剖开的西瓜），模型就会帮我们判断（是不是好瓜）。

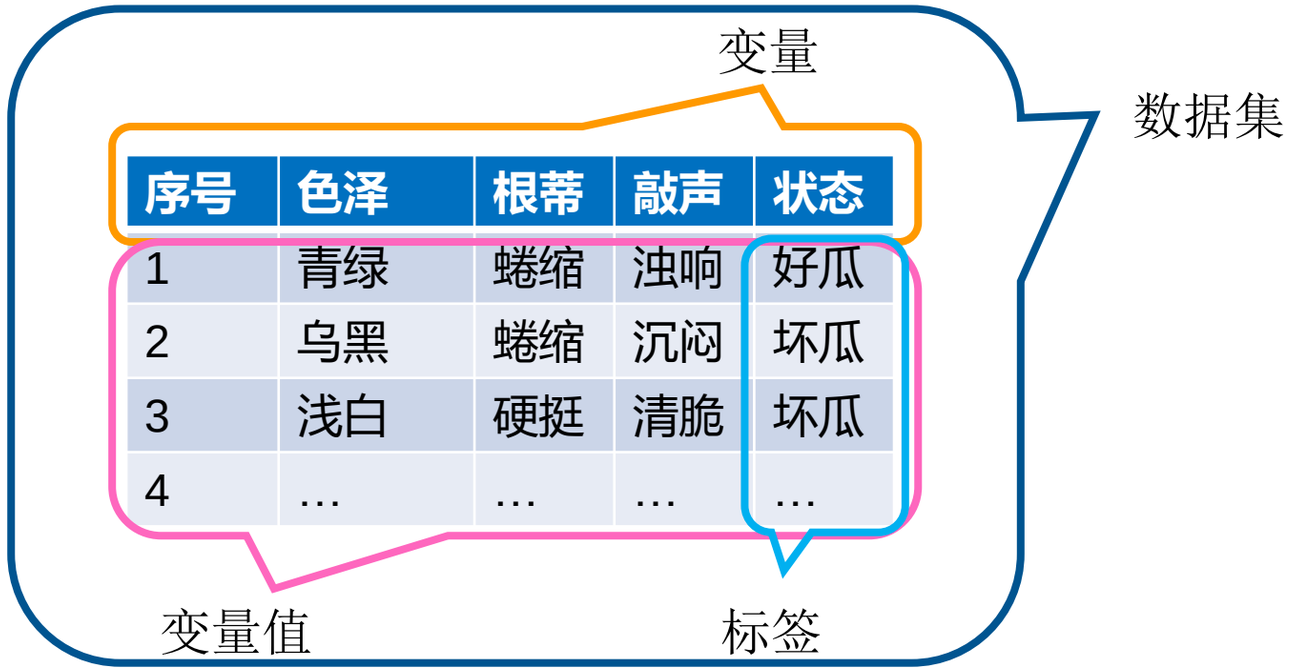


数据挖掘理解

用高中生能理解的数学语言来讲，建模任务的本质就是：

根据一些已有的、从输入空间X（如 {[色泽=青绿；根蒂=蜷缩；敲声=浊响]，[色泽=乌黑；根蒂=蜷缩；敲声=沉闷]，[色泽=浅白；根蒂=硬挺；敲声=清脆]}）到 输出空间Y（如{好瓜，坏瓜，坏瓜}）的对应，找出一个函数f：$X \xrightarrow{f} Y$来描述这个对应关系，这个函数就是我们要的**模型**。

有了模型之后再做**预测**就简单了，也就是拿一套新x，用这个函数算一个y出来就完了。



序号	色泽	根蒂	敲声	状态
1	青绿	蜷缩	浊响	好瓜
2	乌黑	蜷缩	沉闷	坏瓜
3	浅白	硬挺	清脆	坏瓜
4	…	…	…	…

模型 ≠ 函数？

之所以我们更习惯于把模型称为模型而不是函数，因为它不满足我们通常对函数期望的确定性，这里相同的x可能对应不同的y（色泽、根蒂、敲声都相同的瓜可能即有好也有坏）。

数据挖掘理解

模型又是怎么建立出来，也就是这个函数是怎么找出来的呢？

想想如何让一个人拥有判断瓜好坏的能力？需要用一批瓜来练习，获取剖开前的特征（色泽、根蒂、敲声等），然后再剖开它看好坏。久而久之，这个人就能学会用剖开前瓜的特征来判断瓜的好坏了。
朴素地想，用来练习的瓜越多，能够获得的经验也就越丰富，以后的判断也就会越准确。

用机器做数据挖掘是一样的道理，我们需要使用历史数据（用来练习的瓜）来建立模型，而建模过程也被称为**训练**，这些历史数据称为**训练数据集**。

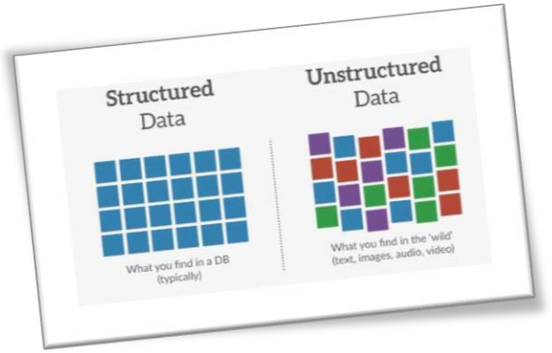
数据挖掘理解

我们通常说要将训练数据先整理成结构化数据才能进行建模，那么什么是结构化数据呢？

结构化数据，即有多种属性构成的批量数据，每条数据（也称记录）的属性（也称字段）相同，一般按行呈现。它可以来自于数据库，也可以来自于文本，或者是HDFS等文件存储系统。

预测titanic幸存者数据见下图：

PassengerI	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S



数据挖掘理解

显然，训练数据集中必须有我们关心的目标（瓜的好坏），也就是那个Y必须有值（用来练习的瓜，其好坏是知道的），称为目标变量。

训练数据集中当然还要有来判断瓜好坏的特征如色泽、根蒂、敲声，这些称为特征变量。
用结构化数据的术语来讲，目标变量和特征变量都是数据的属性或字段。
预测好瓜和titanic数据的目标变量和特征变量如下：

目标变量

色泽	根蒂	敲声	状态
青绿	蜷缩	浊响	好瓜
乌黑	蜷缩	沉闷	坏瓜
浅白	硬挺	清脆	坏瓜
…	…	…	…

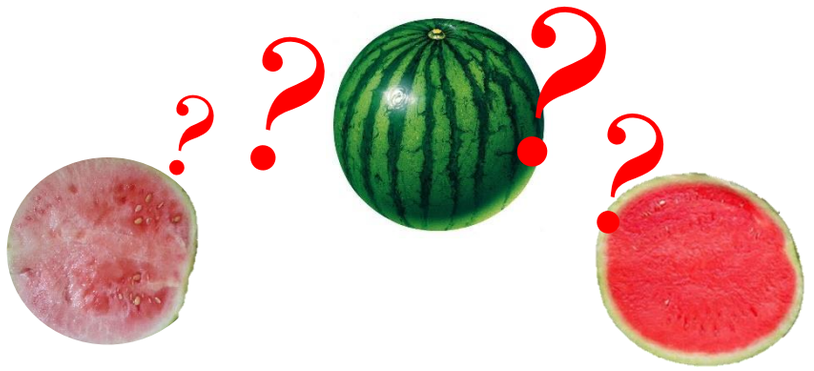
PassengerI	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S

特征变量

数据挖掘理解

需要指出的是，数据挖掘的模型，对新数据的预测不可能做到100%准确，就像再有经验的瓜农也不可能在一个瓜打开之前，100%确定它是好瓜还是坏瓜，只是经验丰富时准确率更高。（其实就前面所说，模型甚至对于历史数据都不可能做到100%准确，同样的特征变量可能对应不同的目标变量）。

既然做不到100%准确，那就存在准确率的概念，即不同模型有不同的准确率。建模的目标是预测，当然准确率越高越好。那么我们就要有办法进行模型评估，以搞清哪个模型的准确率更高。



数据挖掘理解

那么，有什么办法知道模型的准确率呢？

简单，我们把已有目标变量的历史数据分成两部分，一部分用于训练模型，即训练集。另一部分用于评估模型，即测试集。使用训练集训练出来的模型再用测试集来验证，就可以知道模型的准确率好不好。因为这部分数据已经有目标变量了，可以拿已知的目标变量和模型预测出来的目标变量进行比较。

就象考核一个人判断瓜的能力，也要在他判定之后再打开瓜看一下（获取瓜的好坏信息）来看看判定的是不是准。

我最会挑瓜！

从来没有挑到坏瓜！

打眼一看就知好坏！

现场开瓜比赛

数据挖掘理解

分类模型

如果要预测的是离散值，如“好瓜”，“坏瓜”，模型就称为**分类模型**；

如果要预测的是连续值，如西瓜成熟度“0.95”，“0.63”，这个模型称为**回归模型**。

回归模型

分类模型的评估是看分类是否正确，指标有很多（后面会详细介绍）。

回归模型的评估是看预测结果与真实结果的差距，指标也有很多（后面也会详细介绍）。

目录 CONTENTS

02

数据探索

数据探索

建模之前先要针对手里的训练数据进行探索：

- 数据是否达到我们原来设想的要求；
- 样本中有没有什么明显的规律和趋势；
- 有没有出现从未设想过的数据状态；
- 属性之间有什么相关性；
- 它们可区分成怎样一些类别
- ……

用建模的行话说是：

数据质量分析　　数据特征分析

数据探索

首先要对数据有总体认识：有多少个特征，多少条记录，各个特征属于哪些类别……

我们使用kaggle中用于预测tatinic幸存者的经典数据来说明。

用易明智能建模工具打开这个数据文件，可以看到，数据共891条记录，12个变量，其中Survived作为目标变量。

数据探索

常见的数据类型有这些：

变量类型	描述	例1	例2
ID	每条记录的唯一识别标记，这种变量通常没什么用。	乘客编号:[1,2,3,4…]	用户ID:[pdsese32sd,edewdpbnjg43d,…]
单值变量	只包含一个类别的变量（不含缺失值）	家用电压 :[220,220,,…]	是否售出(只记录售出的):[1,,1,1,,,,1…]
二值变量	只有两个类别的变量（不含缺失值），目标变量常常是这种类型的	性别:[男,女]	是否售出:[1,0,1,1,0,0,0,1…]
分类变量	分类数多于两个类别的变量	行业:[旅游业,制造业,IT…]	学科:[数学,语文,英语…]
计数变量	取值为自然数的变量	班级人数:[45,67,53…]	房间数:[2,5,6,7...]
数值变量	取值为实数的变量	身高:[175.5,180.4,165.3…]	销售额:[2300.87,1098,8…]
时间日期	表示时间日期的变量	生日:[2009-01-01…]	登录时间:[2019/1/1 12:00:00,…]
长文本	长度超过128字节且分类数特别多的变量，这种变量一般也没法直接用，需要再做变换	故事简介:[Harry potter says:”……..”,He is ……,……]	

数据探索

Titanic数据的变量情况

序号	变量	描述信息	变量类型
1	PassengerId	乘客编号	ID，唯一识别码
2	Survived	是否幸存	二值变量，目标变量
3	Pclass	船票等级	分类变量
4	Name	乘客姓名	ID，唯一识别码
5	Sex	乘客性别	二值变量
6	Age	乘客年龄	数值变量
7	SibSp	乘客兄弟姐妹、配偶数量	计数变量
8	Parch	乘客父母孩子数量	计数变量
9	Ticket	船票号码	分类变量
10	Fare	船票价格	数值变量
11	Cabin	船舱	分类变量
12	Embarked	登船港口	分类变量

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S

数据探索

不同的数据类型适用于不同的模型，如树模型更适用于分类变量等。需要正确地指定数据类型，并且对不同的类型进行不同的处理。

我们可以使用易明智能建模工具自动识别变量的类型，但有时会有错误，

- 比如当房间数都在20个以下如[2,3,4,8,3,2,1…]，可能被识别成分类变量；
- 再比如当数值变量都是整数时会被识别成计数变量，比如房价[113742,259753,…]；这时需要手动更改

数据探索---数据质量分析

数据质量分析的主要任务是检查原始数据中是否存在脏数据，脏数据包括以下内容：

- 缺失值
- 异常值
- ……

缺失值分析

缺失值影响

1）建模丢失大量信息。
2）增加模型不确定性。
3）建模陷入混乱，不可靠输出。

缺失值分析

1）缺失值的属性有哪些
2）属性的未缺失数
3）缺失数。
4）缺失率。

数据探索---数据质量分析

缺失值分析

对计数变量“Age”进行缺失值分析结果：

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

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

	目标变量为1	目标变量为0	总计
缺失率	15.205%	22.769%	19.865%
平均值	28.344	30.626	29.699
标准差	14.951	14.172	14.526
偏度	0.18	0.584	0.388
最小值	0.42	1.0	0.42
下1/4点	19.0	21.0	20.0
中位数	28.0	28.0	28.0
上1/4点	36.0	39.0	38.0
最大值	80.0	74.0	80.0

分析得知，变量Age的缺失率是19.865%，其中目标变量为1的缺失率为15.205%，目标变量为0的缺失率为22.769%。预处理时，需要对缺失值进行填补。

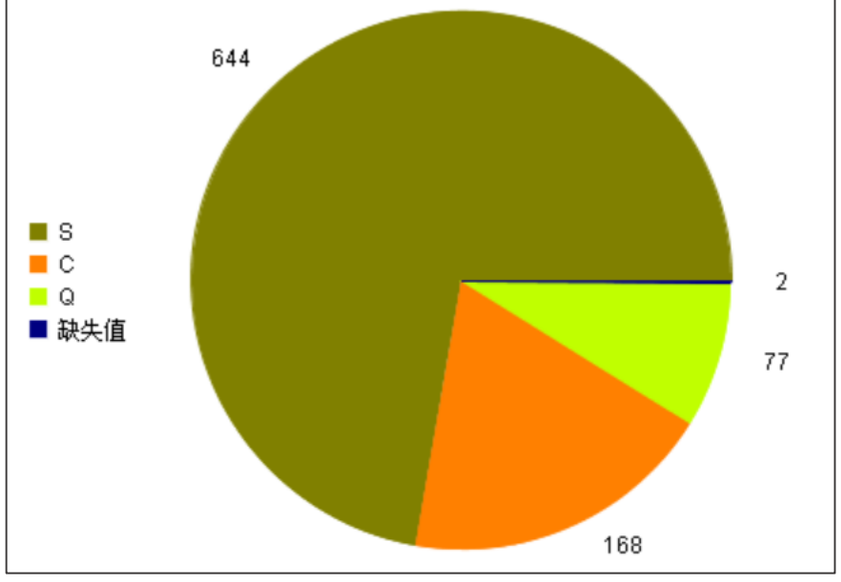
数据探索---数据质量分析

缺失值分析

对分类变量登船港口“Embarked”进行缺失值分析结果：

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%

分析得知，变量“Embarked”的缺失率是0.224%，缺失数是2。预处理时，需要对其进行填补。

数据探索---数据质量分析

异常值分析

异常值是指样本中的个别值，其数值明显偏离其余的观测值。异常值也称为离群点，异常值的分析也称为离群点分析。

箱形图经常用来标记异常值，其中的线和点的含义如下：

中位数：一组数据排序后处于中间位置上的变量值。

上四分位：一组数据排序后处于75%位置的值。

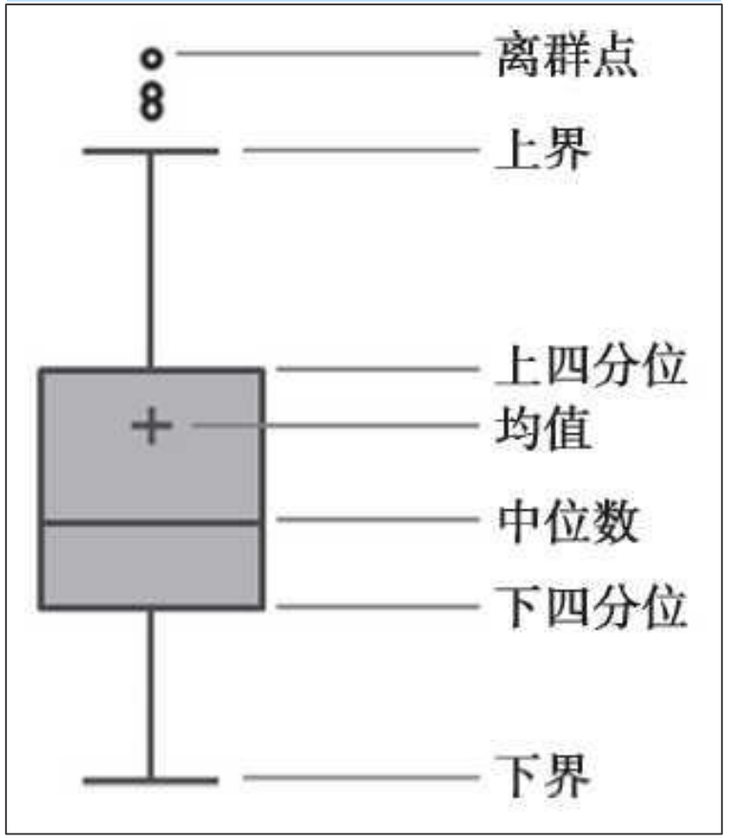
下四分位：一组数据排序后处于25%位置的值。

四分位差：即箱体的长度，上四分位减下四分位。反应中间50%数据的离散程度，数值越小，表示中间数据越集中。

上界：上四分位+1.5*四分位差，大于该值的点作为离群点。

下界：下四分位-1.5*四分位差，小于该值的点作为离群点。

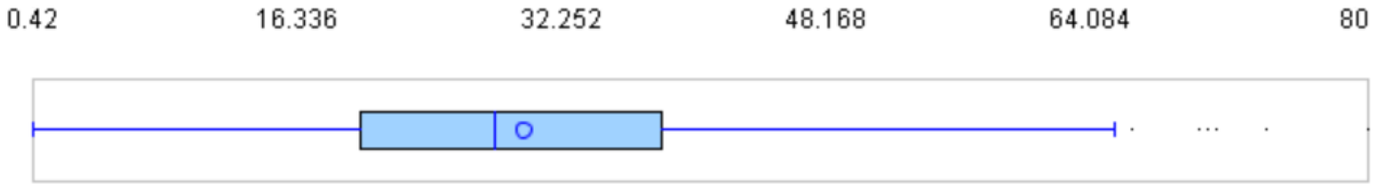
箱形图



数据探索---数据质量分析

缺失值分析

对变量“Age”使用箱形图进行异常值分析：



分析得知，变量“Age”不存在过小的离群点，只有个别几个比较大的离群点，但作为年龄来说，这些值也在可接受范围内，不需要处理。

数据探索---数据特征分析

数据特征分析的内容：

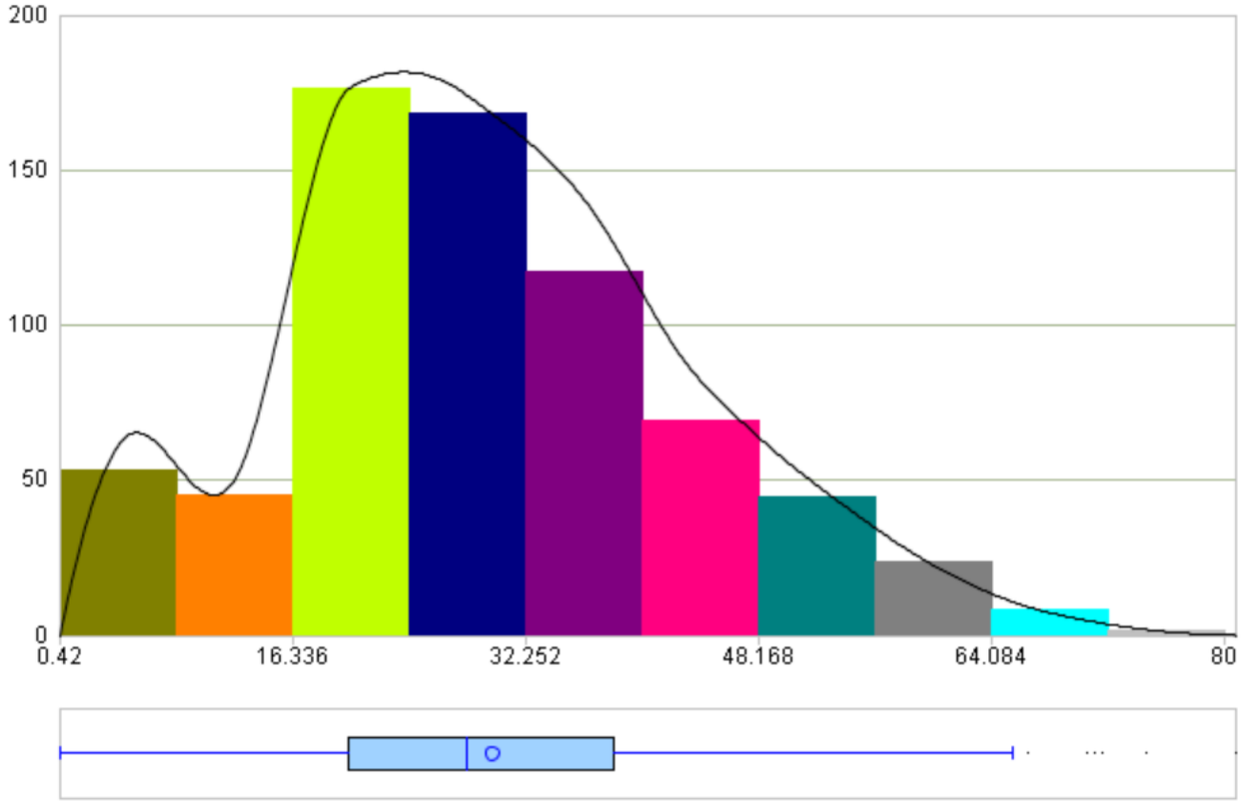
- 分布分析：揭示数据的分布特征和分布类型
- 对比分析：把两个相互联系的指标进行比较
- 统计量分析：用统计指标对定量数据进行统计描述
- 相关性分析：分析变量之间相关程度
- 周期性分析：……
- 贡献度分析：……

……

数据探索---数据特征分析

分布分析---定量数据

对变量“Age”绘制频数分布直方图和箱形图进行分布分析：

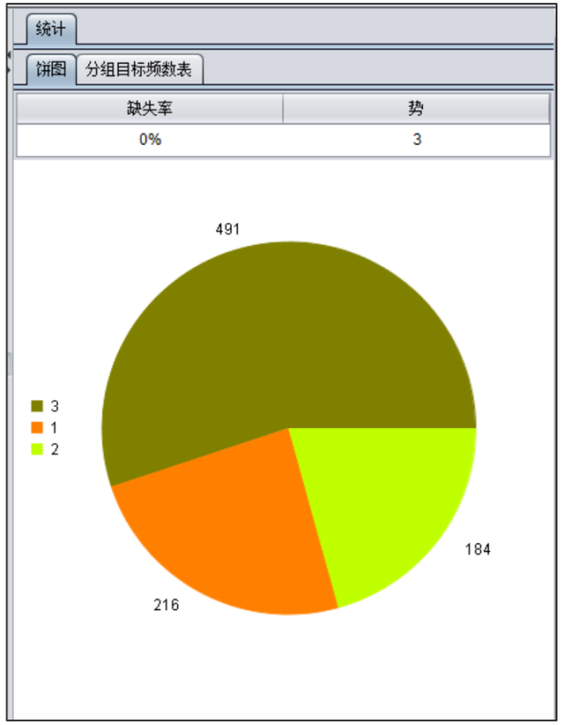


分析发现，年龄在16~48的人数占大多数，近似呈正态分布，微微有偏，只看分布情况是个很不错的变量。

数据探索---数据特征分析

分布分析---定性数据

对分类变量船票等级“Pclass”绘制饼图进行分布分析：



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%

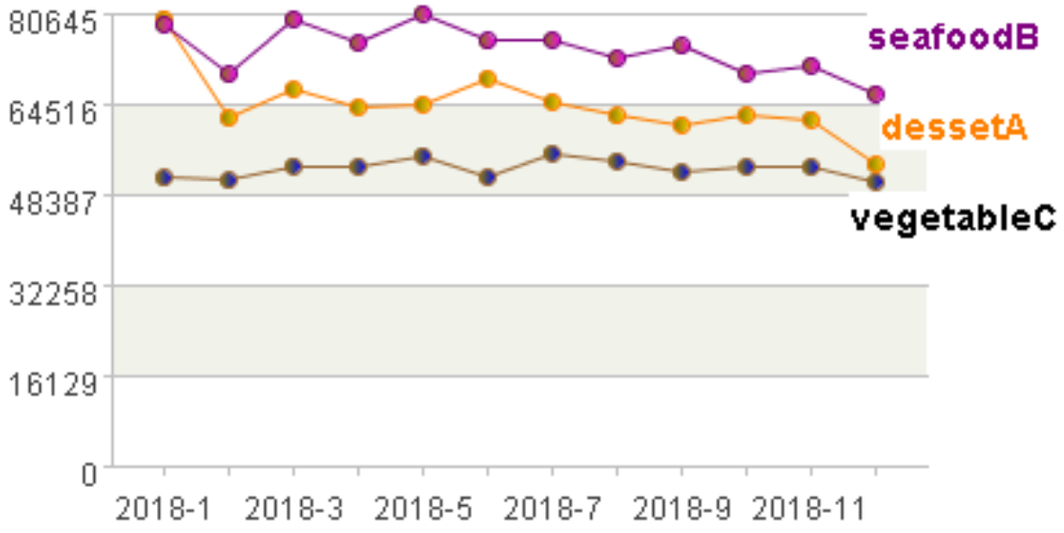
正样本是指那些目标变量取值是我们期望值的记录（比如好瓜、存活），有时也叫阳性数据

分析发现，船上第三等级的乘客多，一二等级的乘客少。按照等级分组后，发现一、二等级的正样本率（幸存率）要大的多，说明该变量对于目标变量的区分度较大，属于重要变量。初步推断：船票等级越高生还率越高。

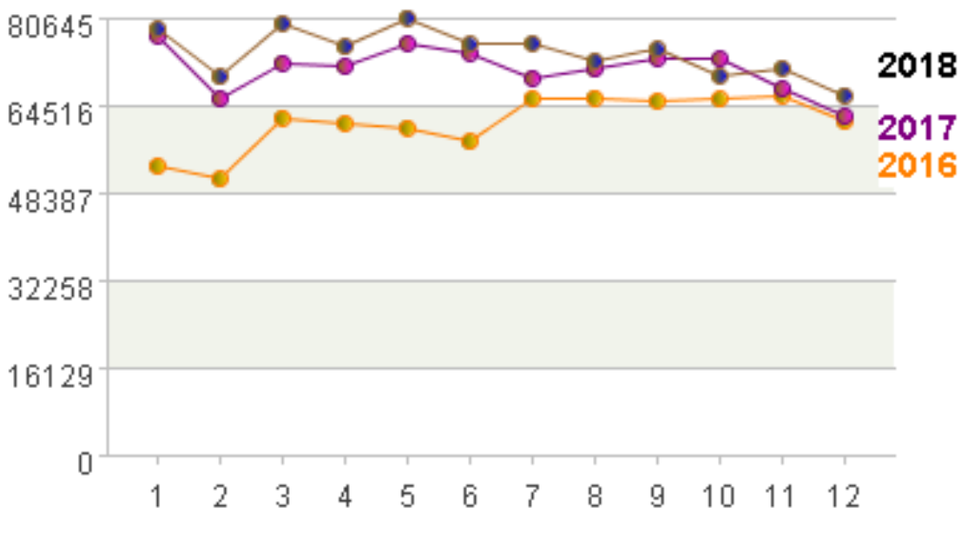
数据探索---数据特征分析

对比分析

对比分析是指把两个相互联系的指标进行比较，从数量上展示和说明研究对象规模的大小，水平的高低，速度的快慢，以及各种关系是否协调。如下图：



2018年某餐饮企业三部门销售额对比图



某餐饮企业海鲜部三年销售额对比图

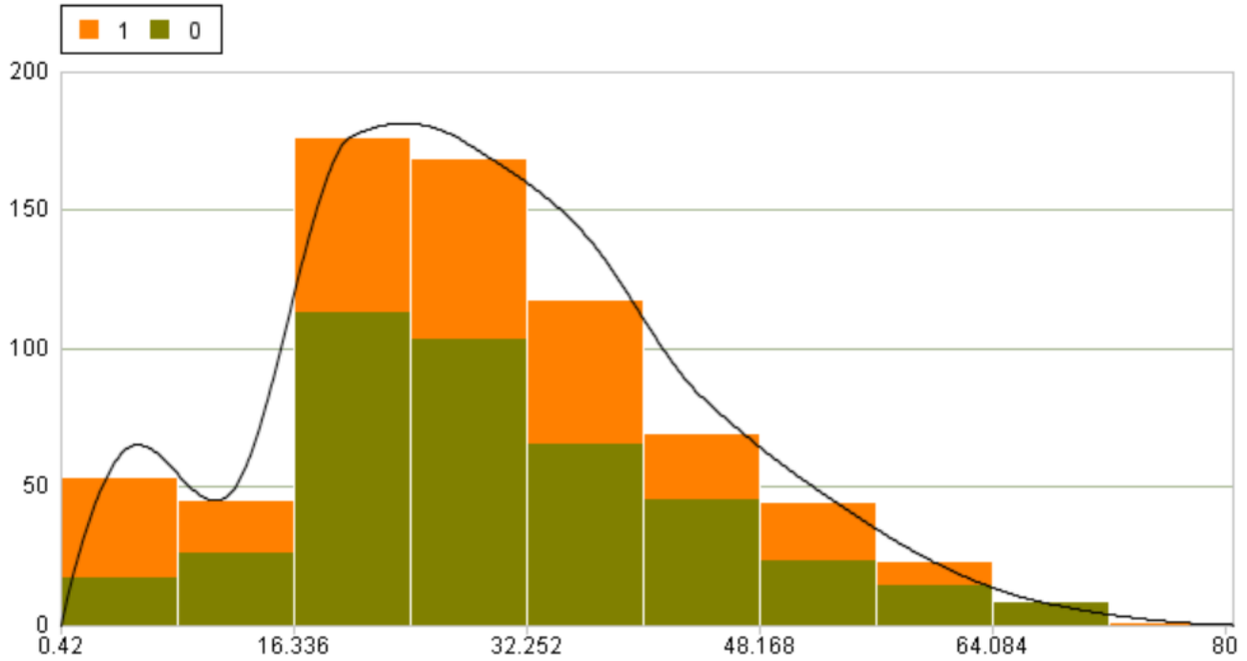
总体看甜品部A和海鲜部B销量呈下降趋势，而素菜部相较于甜品部A和海鲜部B变化平稳。进而可以进一步分析其中原因。

总体看海鲜部B三年的销量是逐年上升的，但17年和18年又在本年度呈现下降趋势，可以从其他的角度分析其中原因。

数据探索---数据特征分析

对比分析

对变量“Age”按照目标变量分组对比分析：



该图不仅是对变量“Age”的分布情况分析，而且还可以对比幸存者（橘黄色1）与遇难者（黄绿色0）在各个年龄段的情况，0~8岁的乘客幸存率很高，超过了一半，而大于56岁的乘客幸存率很低，因此这个变量对目标变量有一定的区分度，是重要变量。初步推断：儿童的生还率高，老年人生还率低。

数据探索---数据特征分析

统计量分析

集中趋势

平均数：一组数据相加后除以数据个数得到的结果。

中位数：一组数据排序后处于中间位置上的变量值。

众数：一组数据中出现次数最多的变量值。

离中趋势

极差：一组数据的最大值与最小值之差。

方差：各变量值与其平均数离差平方的平均数。

标准差：方差开方即得标准差。

数据探索---数据特征分析

统计量分析

对变量船票价格“Fare” 进行统计量分析：

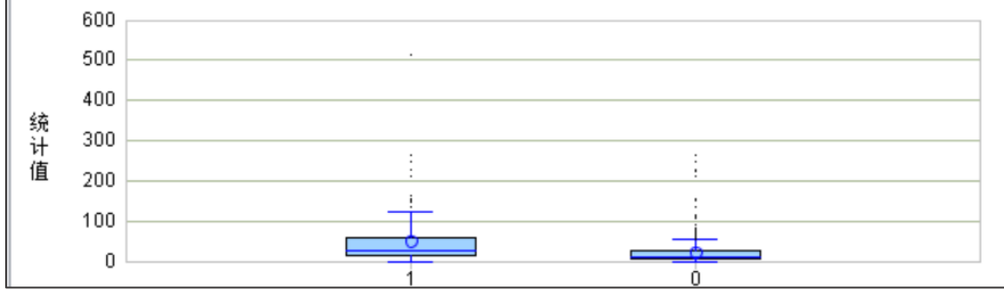
描述性统计量 | 频数分布图 | 分组描述性统计量 | 分组频数分布图

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
0%	0.0	512.329	32.204	31.0	14.454	7.896	49.693	4.779

分析发现，该变量偏度为4.779，需要进行纠偏处理或者进行离散化处理。再看分组后的统计量，目标变量为1和目标变量为0的均值，标准差等差异都较大，说明该变量对目标变量的区分度很大，是个重要变量。初步推断：票价高的乘客幸存率高。

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

	目标变量为1	目标变量为0	总计
缺失率	0%	0%	0%
平均值	48.395	22.118	32.204
标准差	66.597	31.388	49.693
偏度	3.846	4.541	4.779
最小值	0.0	0.0	0.0
下1/4点	12.35	7.854	7.896
中位数	26.0	10.5	14.454
上1/4点	57.0	26.0	31.0
最大值	512.329	263.0	512.329

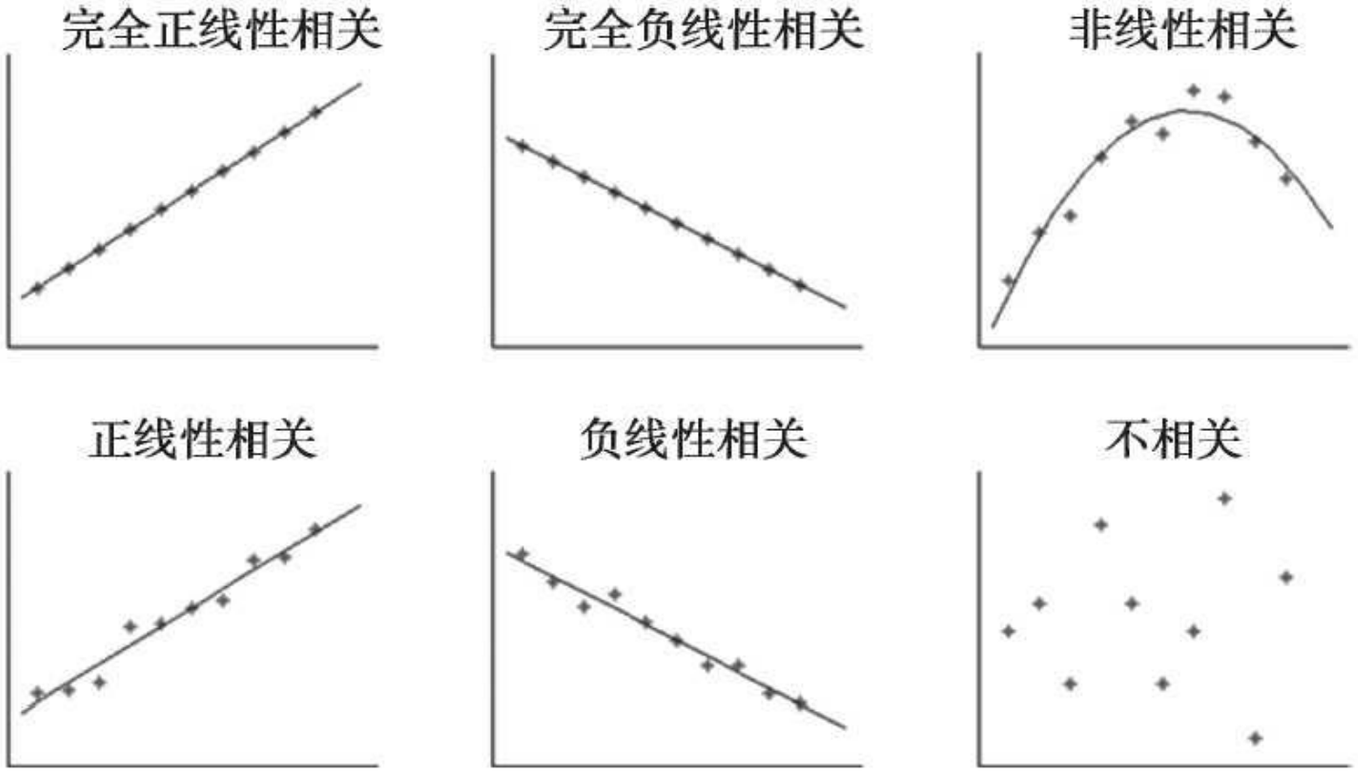


数据探索---数据特征分析

相关性分析

分析连续变量之间线性相关程度的强弱

判断两个变量是否具有线性相关关系的最直观的方法是直接绘制散点图。



数据探索---数据特征分析

相关性分析

可以计算相关系数来更准确地描述变量间的线性相关程度。常用的有Pearson相关系数和Spearman秩相关系数。

1）Pearson相关系数r：用于描述两个连续变量之间的线性相关性；

$$\begin{cases} r > 0, \text{正相关}， \ r < 0, \text{负相关} \\ |r| = 0, \quad \text{不存在线性关系} \\ |r| = 1, \quad \text{完全线性相关} \end{cases}$$

车库面积和房价之间的相关系数

2）Spearman秩相关系数r_s：用于描述两个连续变量之间的等级相关性；

相关系数的绝对值越大，表示两个变量的相关性越大。如车库面积与房价的两个相关系数都大于0.6，表明车库面积越大房价越高。

目录 CONTENTS

02

数据预处理

数据预处理

数据预处理的目的：

- 提高数据的质量；
- 让数据更好地适应特定的挖掘技术或工具

根据数据探索的结果，针对不同的变量类型，进行相应的处理，如对缺失值进行填补、对偏斜严重的变量进行纠偏、数值数据离散化、增加衍生变量、去除冗余变量等。

数据预处理

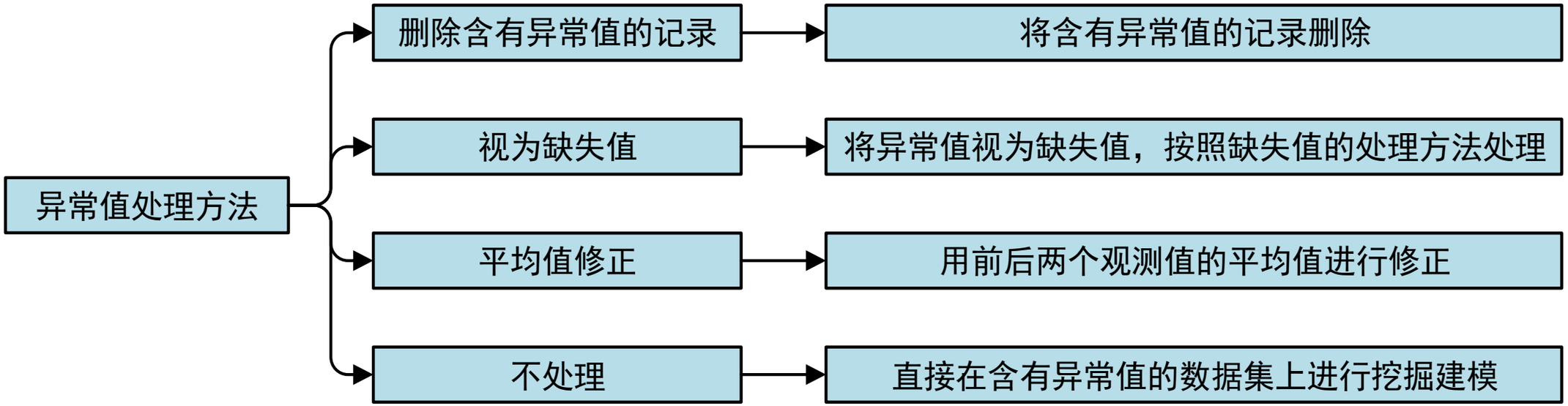
缺失值处理

缺失值处理有三种方法：删除记录、数据插补和不处理。建模时为保证数据信息完整性，通常对缺失值进行插补，常用的插补方法如下：

插补方法	方法描述
均值/中位数/众数插补	根据属性值的类型，用该属性值的平均数/中位数/众数进行插补
使用固定值插补	将缺失的属性值用一个常量替换。如用某地外来务工人员的平均工资来插补某位外来务工人员的“基本工资”。
最近邻插补	在记录中找到与缺失样本最接近的的样本的该属性值插补
回归方法插补	对带有缺失值的变量，根据已有数据和与其有关的其他变量的数据建立拟合模型来预测缺失值的属性值
插值法	插值法是利用已知点建立合适的插值函数f(x)，未知值由对应点xi求出的函数值f(xi)近似代替

数据预处理

异常值处理



在很多情况下，要先分析异常值出现的可能原因，再判断异常值是否应该舍弃，如果是正确的数据，可以直接在具有异常值的数据集上进行挖掘建模。比如之前分析的“Age”变量，虽然用箱形图分析，有少数的比较大的值被认为是离群点，但仍属于正常的范围，所以可以不处理直接进行建模。

数据预处理---数据变换

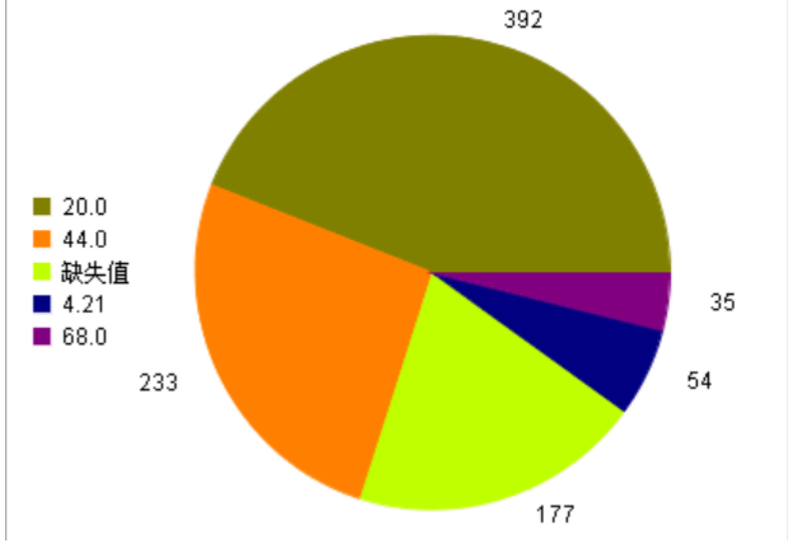
数据变换主要是将数据转换成“适当的”形式，以适用于挖掘任务及算法的需要。

数据变换—离散化

前面数据探索得知变量“Age”是重要变量，其中儿童生还率很高，青壮年的生存率差距不大，中老年人生存率较低，将年龄分为0~8,9~32,33~56,57以上4个分组。如右图：

饼图 | 分组目标频数表

缺失率	势
19.865%	5



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%
4.21	54	36	66.667%
20.0	392	147	37.5%
44.0	233	97	41.631%
68.0	35	10	28.571%

可以发现，离散化后，儿童的那组幸存率高达66.67%，而老年组则只有28.57%，区分度更大，更有利于建模

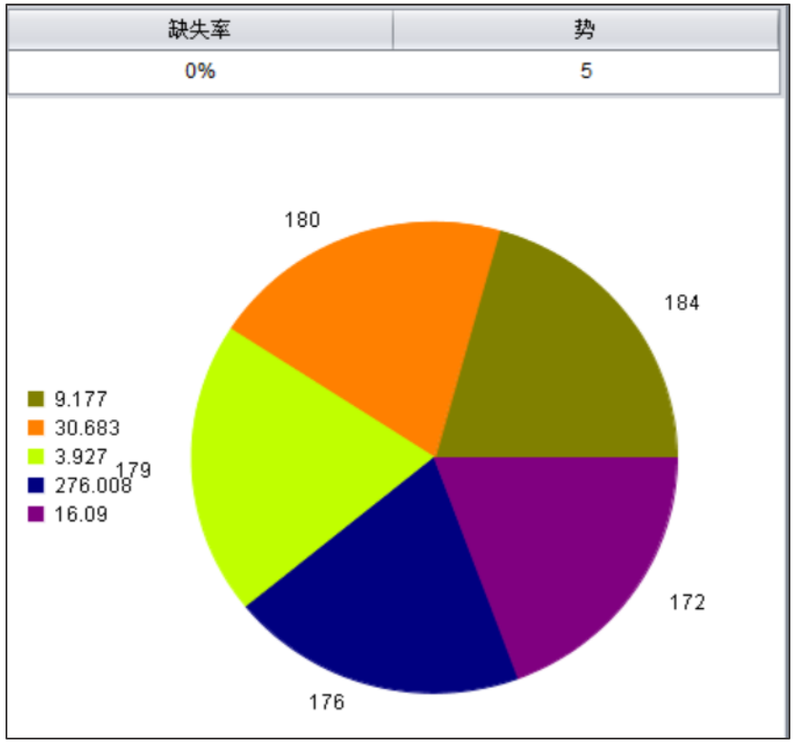
年龄“Age”进行离散化

数据预处理---数据变换

数据变换—离散化

前面数据探索得知变量“Fare”的偏度过大，不适于训练模型，可以将其离散化为分类变量。如下图：

缺失率	势
0%	5



船票价格“Fare” 进行等频离散化

饼图 | 分组目标频数表

分类变量	样本里	正样本数	正样本率
3.9271	179	39	21.788%
9.1771	184	37	20.109%
16.0896	172	73	42.442%
30.68335	180	80	44.444%
276.00835	176	113	64.205%

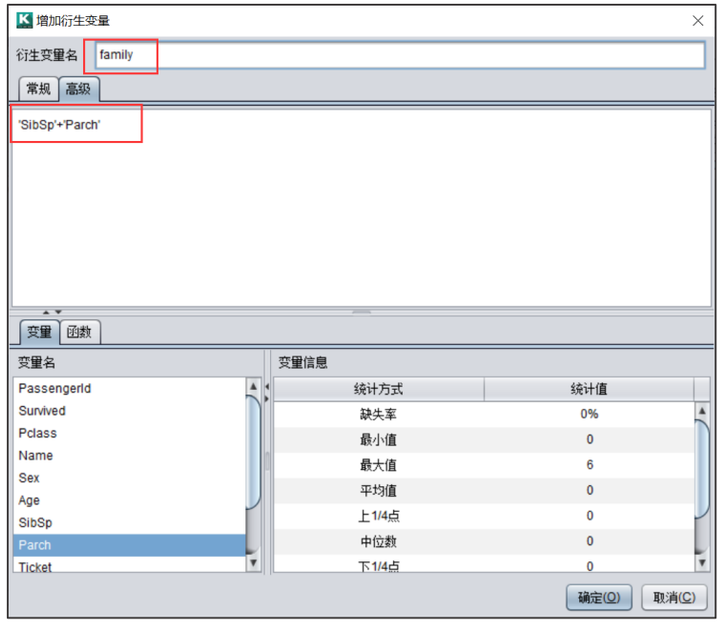
可以发现，离散化后，船票价格越高的分组幸存率越高，价格最高的分组更是高达64.2%，这种区分度高的变量更有利于建模。

数据预处理---数据变换

数据变换—增加衍生

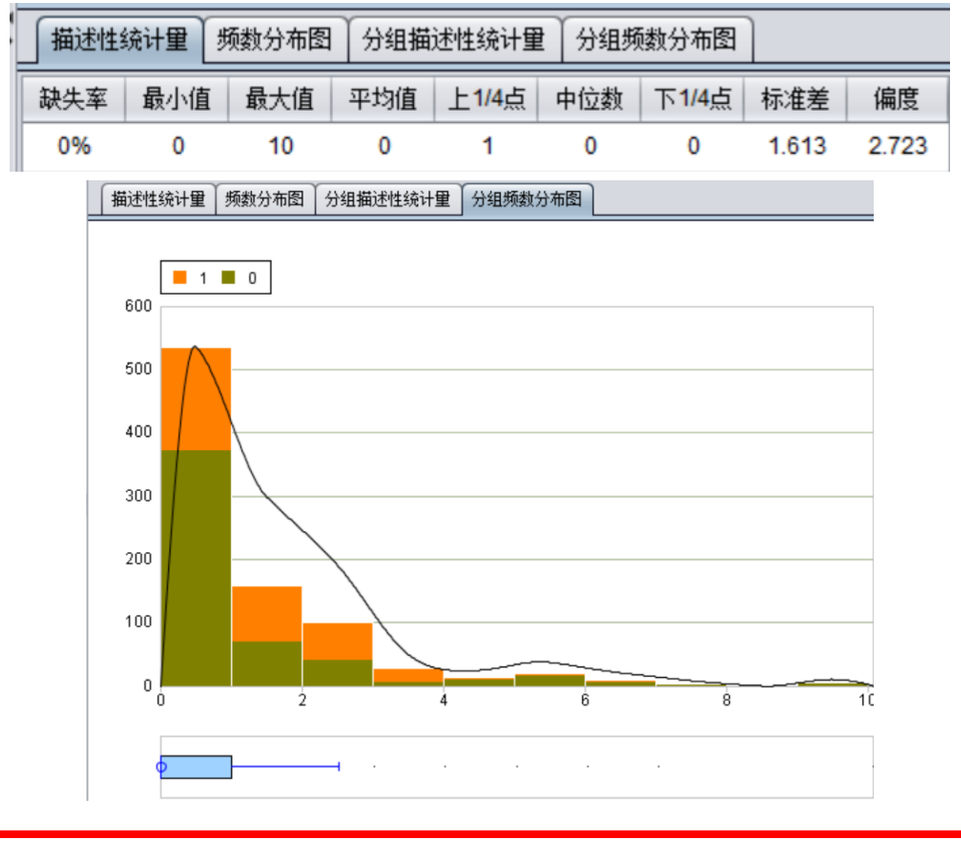
用变量姐妹、配偶数量“SibSp”和 变量父母、子女数量“Parch”相加得到家庭成员数量“family”

增加衍生变量“family”



观察“family”分组频数分布图，可以看到家庭成员数量在1~3人时，乘客的幸存率是比较高的。对建模效果可能会提升。

对衍生变量“family”统计



数据预处理---变量选择

建模前，我们可以去除掉一些无关变量，如PassengerId，Name都是乘客的唯一标识，从直观上就能感觉到它与目标变量无关。
在预处理时我们用SibSp+Parch生成了family
所以SibSp和Parch就可以去掉了。因此我们选出的变量如右图：

目标变量 Survived 设置 筛选变量

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		☑
13	family	分类变量		☑

数据预处理

数据探索和预处理非常重要，但其过程已经有些专业和艰深，需要相当的统计学基础。不过，我们可以使用易明智能建模工具自动实施。

右图是易明智能建模给出的数据预处理报告。其中包含了缺失值处理、数据纠偏、数据变换等数据预处理方法。为建模提供了高质量的建模数据。

变量处理信息

变量名称：，变量类型：ID
变量名称：Pclass，变量类型：分类型变量
分类个数：3
该变量采用简单方式填补缺失值：3。
生成分类衍生变量：BI_Pclass_1, BI_Pclass_2
变量名称：Sex，变量类型：二值型变量
分类个数：2
该变量采用简单方式填补缺失值：male。
生成分类衍生变量：BI_Sex_1
变量名称：Age，变量类型：数值型变量
偏度：0；　　平均值：29.699；
中位数：24；　方差：13.002。
该变量采用简单方式填补缺失值：29.69911764705882。
变量名称：SibSp，变量类型：计数型变量
偏度：0；　　平均值：0.523；
中位数：0；　方差：1.103。
该变量采用简单方式填补缺失值：0。

目录 CONTENTS

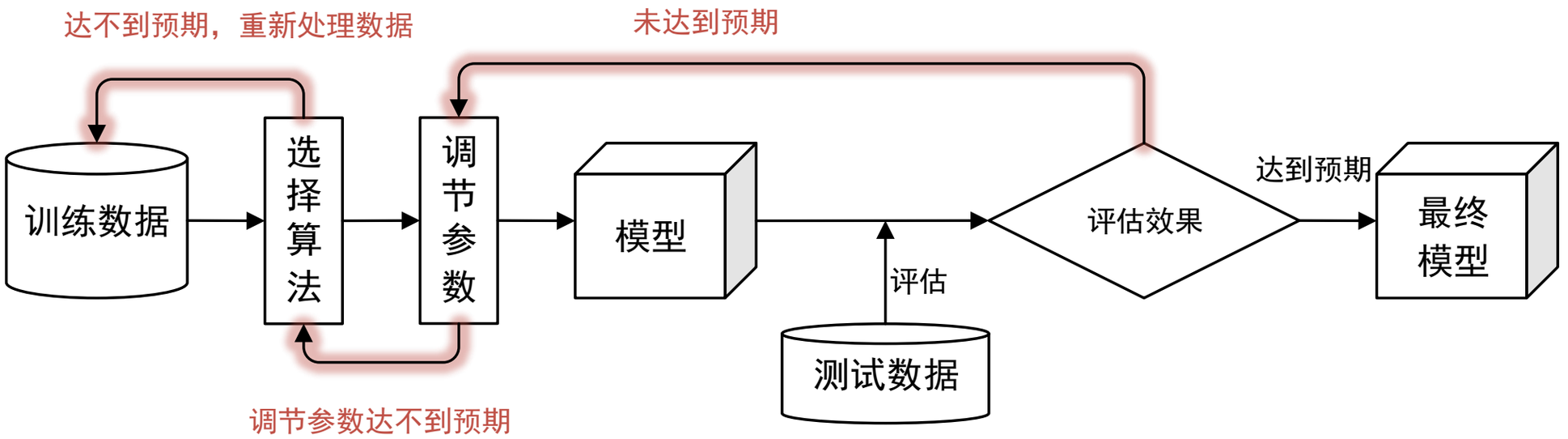
04

挖掘建模

挖掘建模

如前所述，为了验证模型，需要将数据分成训练集和测试集两部分。
这个切分最好是随机的，即训练集和测试集不会受顺序的影响，常用的比例是训练集70%、测试集30%。

训练集准备好后可以进行建模。建模是一个复杂的过程，示意图如下：



挖掘建模---算法选择

好多算法既可以用于分类也可以用于回归，常用的算法及适用数据如下表：

算法名称	适用数据
逻辑回归（Logistic）	特征变量较少的数据集
线性回归（Ridge）	线性数据集
决策树(DT)	分类变量多但每个变量的类别数少的数据集
随机森林（RF）	分类变量多但每个变量的类别数少的数据集
梯度提升迭代决策树（GBDT）	特征变量不是特别多的数据集
极端梯度提升（XGB）	是一个比较全面的算法
提示： 按照算法的优度排序大致是这样： XGB>=GBDT>=RF>=DT>=Logistic>=线性回归（仅供参考哈） 智能建模是从以上的算法中选出最佳模型进行组合，从而得到最佳效果。	

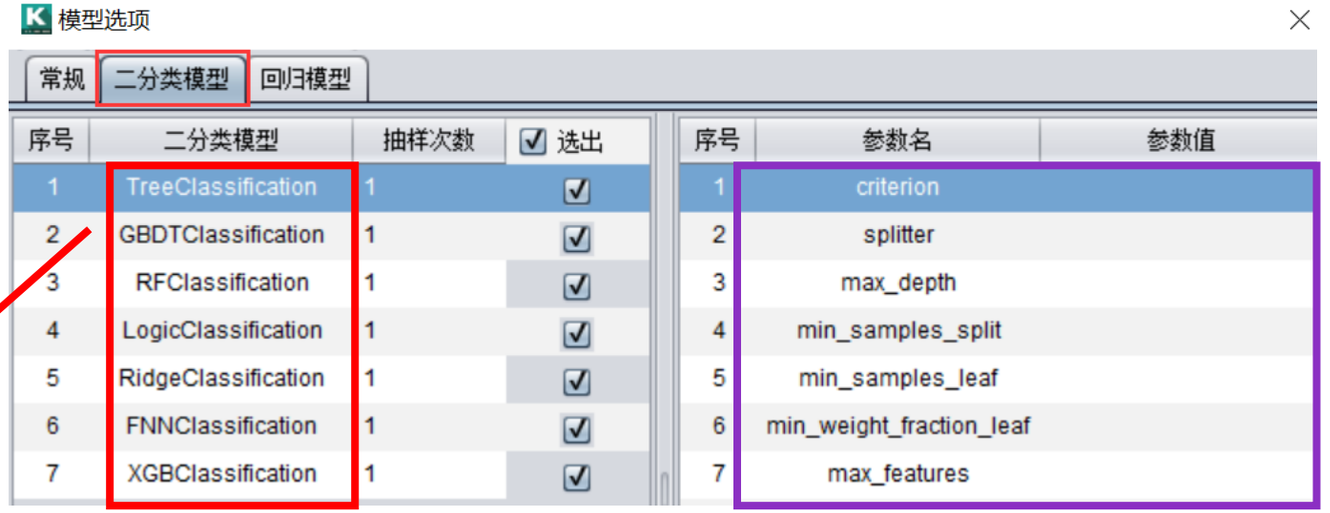
挖掘建模---建模工具

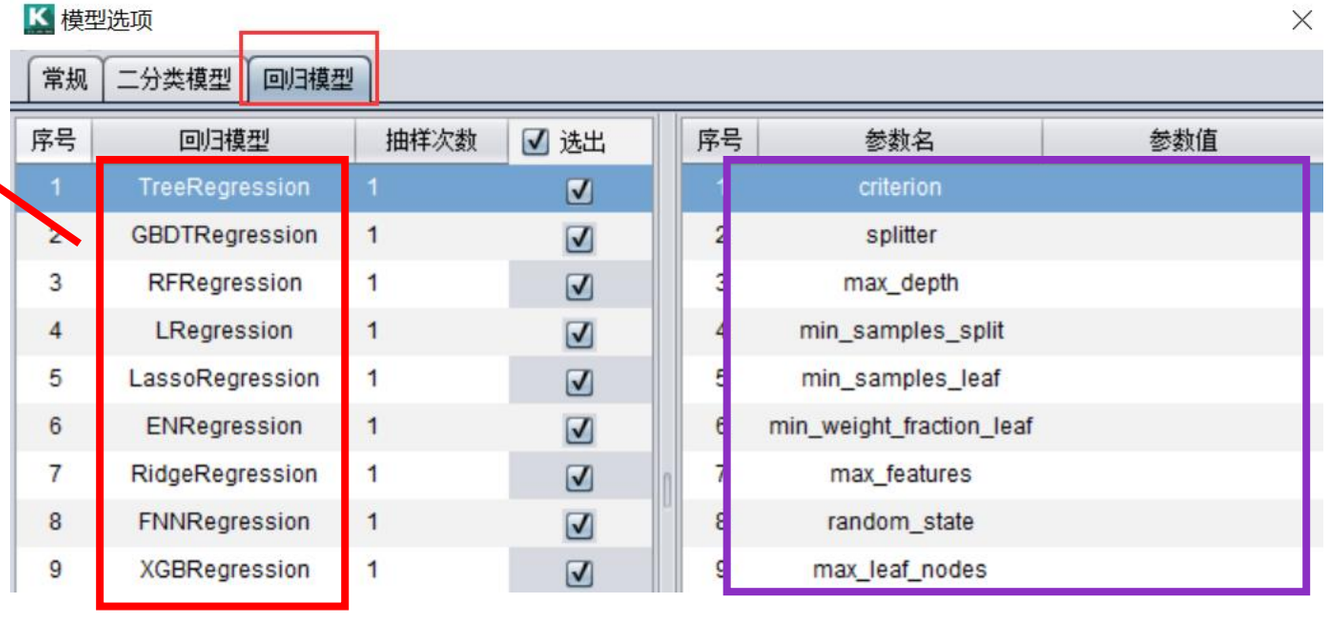
模型选择及参数调整是一个非常复杂且专业的过程，这里我们使用易明智能建模工具来完成建模

分类模型

回归模型

智能建模工具可以自动智能的完成模型选择和参数调整，不需要人工参与。数据挖掘专家可以根据需求手动设置参数调整模型





参数设置

挖掘建模---模型评价

分类模型

分类模型的预测结果是一个概率，表示是正样本的可能性。

- 如使用模型对西瓜进行预测，得到的结果是0.8，则表示该西瓜有80%的可能是好瓜。结果是0.2，则表示只有20%的可能是好瓜，80%的可能是个坏瓜。

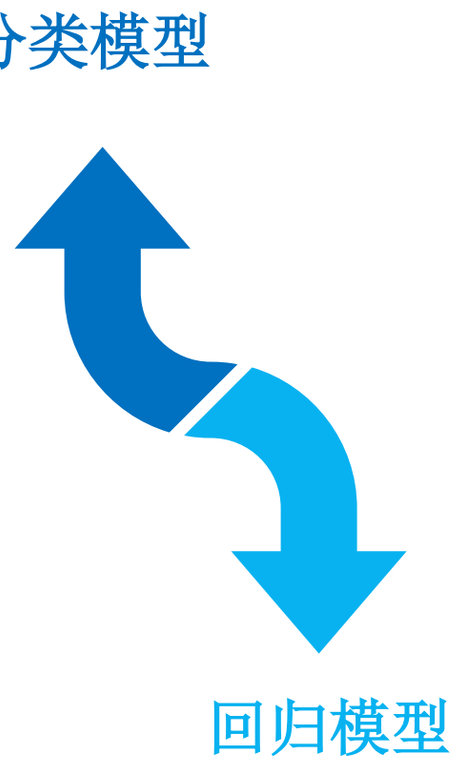
因此我们需要设置一个阈值h来区分好瓜和坏瓜。

- 如h>=0.6则是好瓜，h<0.6则是坏瓜。使用不同的阈值得到的好瓜数自然也不同。

回归模型

回归模型的预测结果是目标变量的一个预测值。

- 如预测房价为55W，表示该房子的价格可能是55W，但只是个参考值，而不能作为最终的结论。



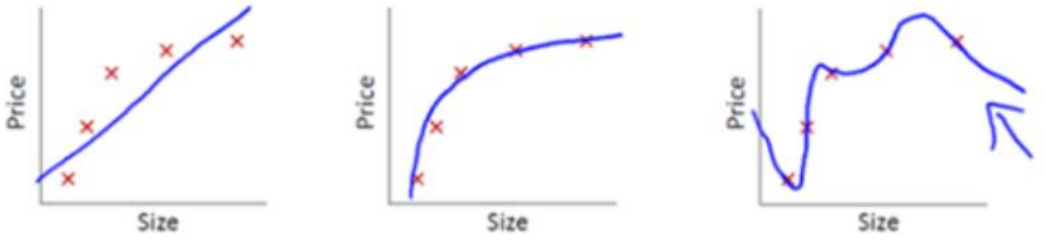
挖掘建模---过拟合

模型要避免过拟合，什么是过拟合？它又是如何产生的

使用训练集建模，建出的模型首先要在训练集上表现够好（准确率高），事实上我们有一些数学方法能做到这一点。但如果模型在测试集上表现不好（准确率很低），那这个模型也是没有用的。这种在训练集上表现特别好但测试集上效果差的情况就称为过拟合，通常是过度使用训练集及相关的数学方法导致的。

在训练集和测试集上表现相差不大的模型才是好模型。

右图是房子大小与房价之间的拟合模型，我们希望通过拟合出一条曲线来预测房价。直观上看，第二张图的曲线（模型）最合适，图一是欠拟合，图三过拟合。



欠拟合

过拟合

挖掘建模---模型评价

如何评价模型？

分类模型评价

- **常用指标**：准确率，精确率，查全率，AUC，GINI，KS等
- **可视化**：ROC曲线，提升度图，查全率图

回归模型评价

- **常用指标**：R^2，MSE，RMSE，GINI，MAE，MAPE等
- **可视化**：残差图，结果对照图

模型评价---混淆矩阵

混淆矩阵是对分类模型预测结果的总结，使用矩阵来汇总预测正确和不正确的数量。**它显示了分类模型进行预测时会对哪种情况产生混淆**。基于混淆矩阵可以得出模型的其它表现评估指标。

下表是在测试集（10个已知好坏的西瓜）上得到的一个混淆矩阵：

真正类数（TP）=3, 实际是好瓜，预测也是好瓜的数量

假正类数（FP）=1, 实际是坏瓜，预测是好瓜的数量

假负类数（FN）=2, 实际是好瓜，预测是坏瓜的数量

真负类数（TN）=4, 实际是坏瓜，预测也是坏瓜的数量

		预测值		
		好瓜	坏瓜	总数
真实值	好瓜	3，TP	2，FN	5
	坏瓜	1，FP	4，TN	5
	总数	4	6	10，N

准确率（ACC）=(TP+TN)/N, 所有预测结果有多少是正确的，也就是好瓜预测成好瓜、坏瓜预测成坏瓜

精确率（PPV）=TP/(TP+FP), 预测成好瓜的结果中，有多少真是好瓜

灵敏度（TPR）=TP/(TP+FN), 实际的好瓜中，有多少被正确预测了

特异度（TNR）=TN/(FP+TN),实际的坏瓜中，有多少被正确预测了

模型评价---混淆矩阵

为什么搞这么多指标，有一个准确率不就够了吗？　　嗯，并不够。

不同的场景目标需要不同的评估指标。比如精确率可以这样用：某企业希望销售50件产品，该企业建立了两个模型来选择待推销客户，混淆矩阵如下：

模型A		预测值		
		买	不买	总数
真实值	买	100，TP	20，FN	120
	不买	50，FP	70，TN	120
	总数	150	100	240，N
精确率=100/150=0.667		准确率=(100+70)/240=0.708		

模型B		预测值		
		买	不买	总数
真实值	买	50，TP	70，FN	120
	不买	10，FP	110，TN	120
	总数	60	180	240，N
精确率=50/60=0.833		准确率=(50+110)/240=0.667		

只考虑准确率，似乎应当选择A模型，但这时候我们需要对75（=50/0.667，预测购买者中有66.7%的实际会购买，即精确率）个客户推销才可能卖出50件商品；而选择模型B，则只要对60（=50/0.833）个客户推销就可能卖出50件商品了，推销成本反而降低了。

因为，我们只关心能被推销成功的那些客户，而不能成功推销且被正确预测为不能成功推销的，虽然有助于提高模型的准确率，对我们却没什么意义。这个场景下，应当用精确率来评估模型的好坏。

模型评价---混淆矩阵

再看灵敏度作为评估指标的情况，常常用于数据分布不平衡时。
比如机场识别恐怖分子，因为恐怖分子是极少数，如果使用准确率来评估模型的话，那只要把所有人都识别成正常人，其准确率可以达到99.999%甚至更高，但显然这种模型没什么用，这时就需要建立一个灵敏率高的模型，比如两个混淆矩阵如下：

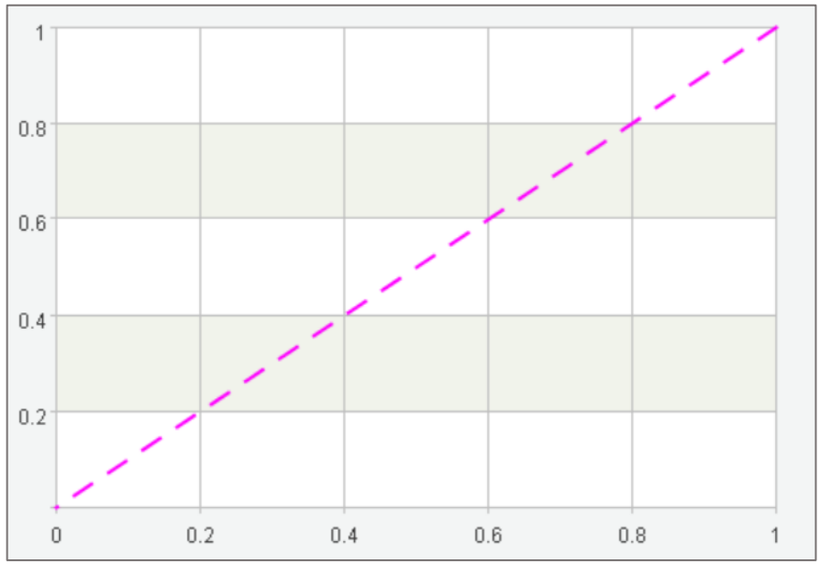
模型A		预测值		
		恐怖分子	正常人	总数
真实值	恐怖分子	0	5	5
	正常人	0	999995	999995
	总数	0	1000000	1000000
灵敏度=0/5=0		准确率=99.9995%		

模型B		预测值		
		恐怖分子	正常人	总数
真实值	恐怖分子	5	0	5
	正常人	95	999900	999995
	总数	100	999900	1000000
灵敏度=5/5=1		准确率=99.9905%		

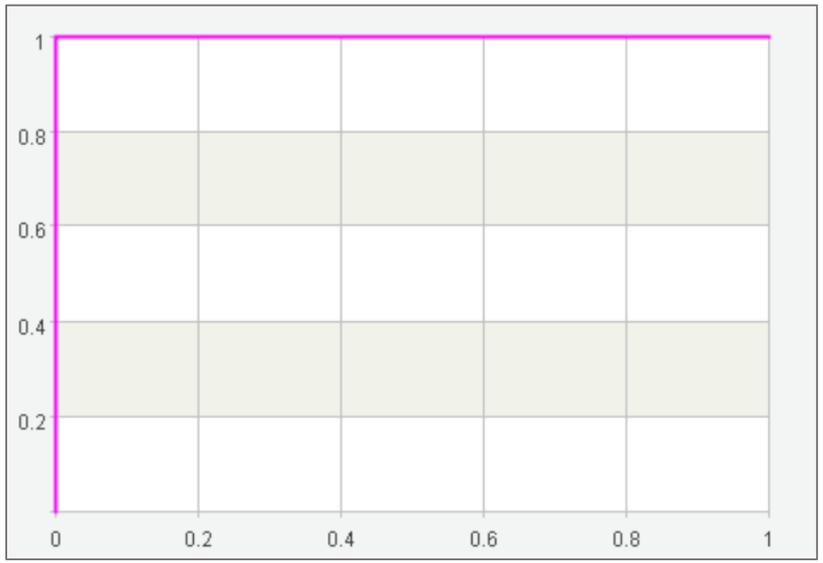
只考虑准确率，会选择A模型，但它根本无法识别恐怖分子。而模型B，虽然准确率低，但可以把全部恐怖分子都识别出来，尽管可能冤枉几个好人，但总比被恐怖分子钻空子好的多。

分类模型评价--- ROC、AUC

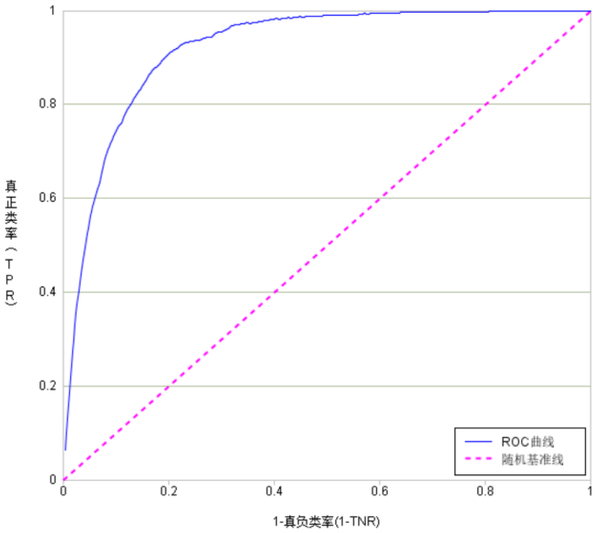
ROC 曲线是评价模型好坏的可视化图形，其上的点由混淆矩阵计算得到的，横坐标是“1-特异度”，纵坐标是“灵敏度”。AUC就是曲线下的面积（Area Under Curve，简单吧），是评估模型总体效果的统计量，AUC越大表示模型越好。



随机模型
AUC=0.5
随机猜测结果
AUC<0.5说明这模型建得还不如抛硬币来得好



完美模型
AUC=1
预测全部正确
但真要AUC=1，那多半是过拟合了

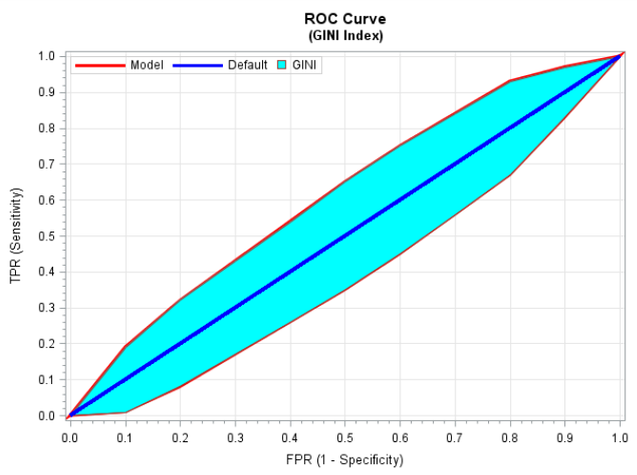


正常ROC曲线0.5<AUC<1
找到一些数据规律，也没有过拟合
这是基本可用的模型

分类模型评价--- Gini指数

Gini指数在数值上等于2*AUC-1（就是那个用来衡量收入不平等的Gini系数，值越大表示贫富差距越大）。在模型评价时，用于表征模型对正负样本（好瓜坏瓜）的区分能力。

如右侧两表所示，模型二的Gini更大，排序后正负样本可以更好的区分，即区分能力更强。



Gini图形效果

模型一预测结果表	
真实值	预测概率
1	0.95
1	0.85
1	0.75
0	0.65
0	0.55
1	0.45
1	0.35
0	0.25
0	0.15
0	0.05
AUC=0.84	Gini=0.68

模型二预测结果表	
真实值	预测概率
1	0.90
1	0.80
1	0.70
1	0.65
0	0.55
1	0.45
0	0.35
0	0.20
0	0.10
0	0.05
AUC=0.96	Gini=0.92

分类模型评价--- 提升度

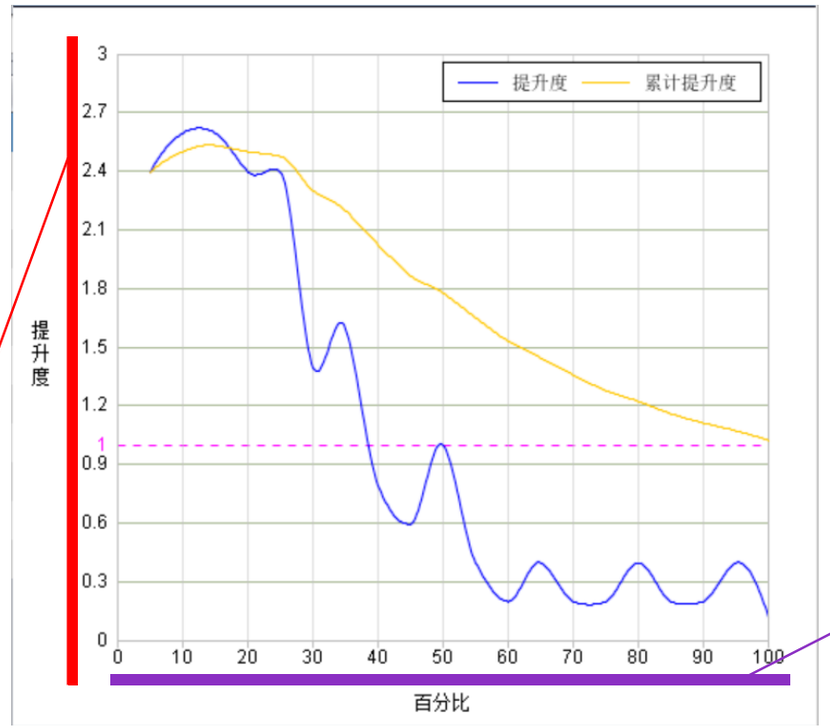
提升度是对模型的有效性度量，它的值为使用和不使用预测模型获得的结果之间的比率。是对模型精确率的评估指标。

假设有100个西瓜，其中好瓜有50个，坏瓜有50个，好瓜率为50%。使用模型对这批西瓜进行预测，按照预测的概率降序排列，前10个瓜中有8个确实是好瓜，预测正确的比例为0.8，则该模型在前10%的瓜中提升度为0.8 / 0.5 = 1.6，0.8是使用模型的正确率，0.5是随机抓的正确率，两者之比即为提升度。也就是说，对于用模型抓出来的前10%的瓜中，使用模型会比随机抓好了多少倍。

好瓜率	前X%	西瓜数	累计样本数	好瓜数	好瓜率	累计好瓜数	累计好瓜率	提升度	累计提升度
0.5	10%	10	10	8	0.8	8	0.8	1.6	1.6
	20%	10	20	7	0.7	15	0.75	1.4	1.5
	30%	10	30	6	0.6	21	0.7	1.2	1.4
	40%	10	40	6	0.6	27	0.675	1.2	1.35
	50%	10	50	5	0.5	32	0.64	1	1.28
	……								

分类模型评价--- 提升度图

提升度特别适合有针对性的营销场景，比如对模型预测概率最高的前10%客户进行促销，成功率会比随机寻找10%客户更高，提升的倍数就是模型的提升度了。好的模型，可能提升数倍到数十倍，这样可以大幅提升促销效率，减少无效推销动作。



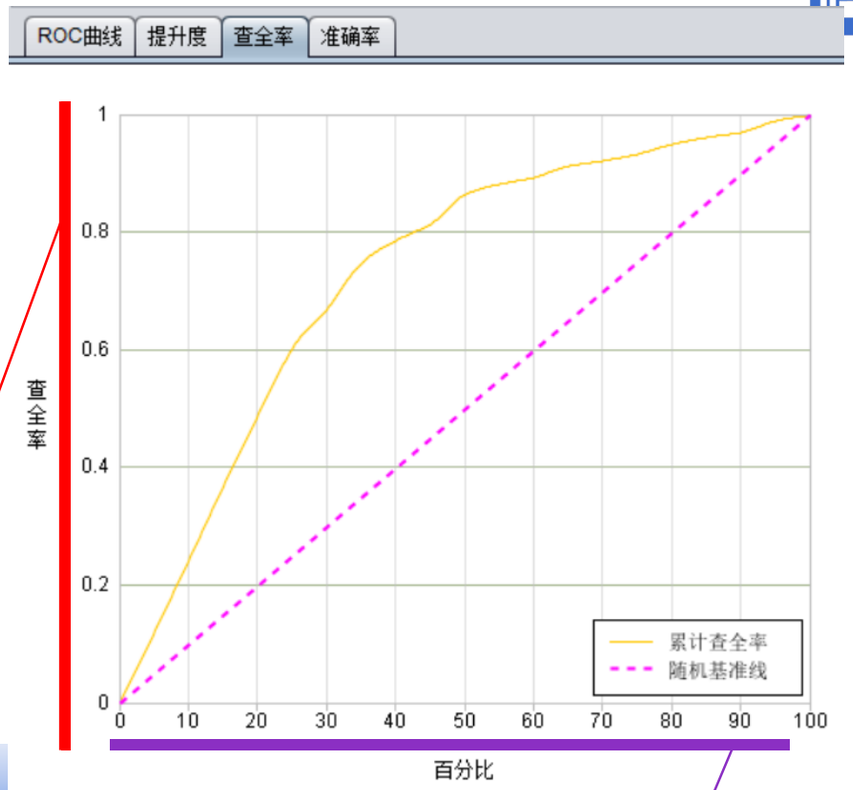
纵轴：提升度

横轴：分组情况

提升度示意图

分类模型评价--- 查全率图

查全率图显示模型找到正类样本（好瓜）的情况，主要应用在数据不平衡的场景。例如，100个西瓜中只有5个是好瓜，95个坏瓜，即使全都预测成坏瓜，这个模型准确率也很好（高达95%），但显然没什么意义。而此时就可以用查全率图来评价模型好坏，斜率越大表示寻找好瓜（术语称为召回率或查全率）的速度越快，模型越好。



查全率示意图

纵轴：累计查全率

横轴：分组情况

累计查全率是各组累计正类数与总正类数的比值。计算过程如下表：

好瓜数	前x%	西瓜数	累计样本数	好瓜数	累计好瓜数	累计查全率
5	10%	10	10	2	2	0.4
	20%	10	20	1	3	0.6
	30%	10	30	0	3	0.6
	.....					

回归模型评价--- 常用指标

常用的回归模型评价指标有：MSE, RMSE, MAE, MAPE和R^2

评价指标	描述信息
MSE(均方误差)	预测值与真实值偏差的平方和的平均数
RMSE(均方根误差)	MSE的平方根。数量级与真实值相同，比如RMSE=10，可以认为回归效果相比真实值平均相差10
MAE(平均绝对误差)	预测值与真实值偏差的绝对值的平均数。
MAPE(平均绝对百分比误差)	预测值与真实值偏差比真实值的绝对值的平均数。 MAPE不仅考虑了预测值与真实值的误差，还考虑了误差与真实值之间的比例，在某些场景下，比如房价在50w到1000w之间，100W预测成50W与1000W预测成950W的差距是非常大的。MAPE可以很好的评价模型好坏
这四个标准的范围都是[0,+∞)，当预测值与真实值完全吻合时等于0，即完美模型；误差越大，这些值越大，模型越不好。	

注意：这些指标都是在测试集上计算的。相同的数据集建立的不同的模型之间可以比较这些指标，数据集不同，这些指标则没有可比性。比如A数据集是预测房价的(50W~1000W之间)，B数据集是预测衣服的售价(50~1000)，很显然A数据集预测的误差一定高于B数据集，但不能说A数据集建立的模型不如B数据集。

回归模型评价---R^2

R^2是预测值与观测值的误差平方和与观测值和观测均值之差的平方和的比值。它可以衡量模型与数据的拟合程度，从而得知模型的整体性能。

R^2 的范围是0到1。
R^2 =0.9表示模型解释了90%的不确定性，模型很不错。

2015年，柴静同学发布《苍穹之下》，其中想说明PM2.5和死亡率的相关性，但视频中不慎出现r^2=0.19…字样，这就闹笑话了（模型好差☺ ）

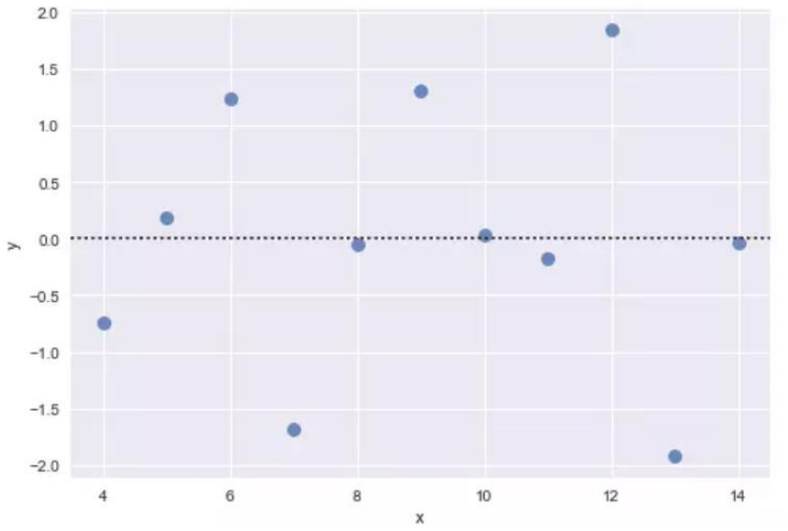
使用相同数据建模，测试集R^2越接近1， 解释不确定性的能力越强，模型越好。
不同的数据集建模的R^2之间没有可比性。

R^2	MSE	RMSE	GINI	MAE	MAPE
0.934573	292043301.1064...	17089.274446	0.197716	12333.645912	7.687109

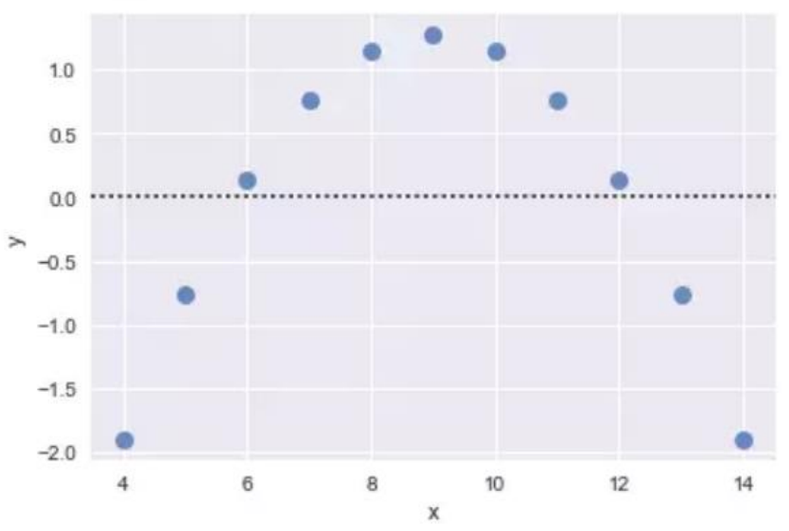
回归模型评价---残差图

残差是观测的Y_i值与模型预测的$f(X_i)$值之差。$\{r_i = f(X_i) - Y_i | \, i = 1, 2, \cdots, N\}$.
残差图是指以残差为纵坐标，以任何其他指定的量为横坐标的散点图。

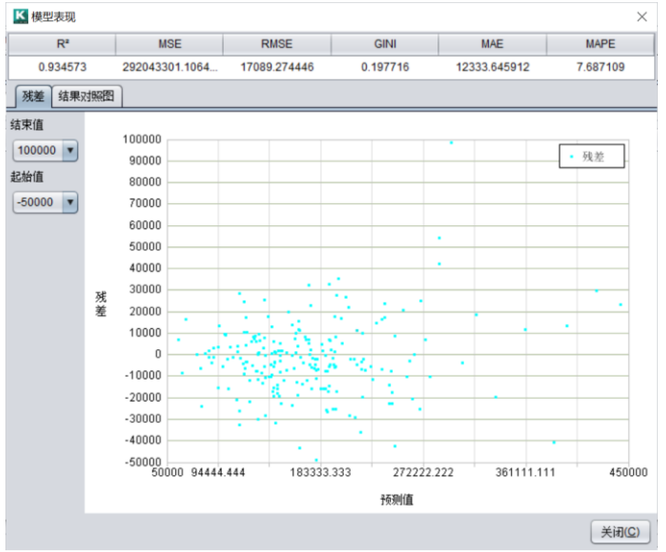
如果残差以对称的模式，随机的散布在0周围，则该模型比较理想。如下图一，
如果可以根据残差图来猜测残差值的大小则表明模型不是理想的。如下图二



图一：理想模型的残差图



图二：不理想模型的残差图



房价预测残差图
不太理想，但可接受

模型评价---变量重要度

建立模型后，还可以得到本次建模时各变量的重要度信息。下图是titanic幸存者的变量重要度信息：

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	Sex	二值变量		☑	1
2	Age	数值变量		☑	0.437
3	Pclass	分类变量		☑	0.338
4	family	分类变量		☑	0.287
5	Cabin	分类变量		☑	0.175
6	Fare	数值变量		☑	0.14
7	Embarked	分类变量		☑	0.051
8	Passeng...	ID		☐	0
9	Survived	二值变量		☑	-
10	Name	分类变量		☐	0
11	SibSp	计数变量		☐	0
12	Parch	计数变量		☐	0
13	Ticket	分类变量		☑	0

	变量重要的作用
1	参考变量重要度，有针对性的对数据重新处理。
2	使用重要度高的变量进行交互生成衍生变量，如路程/时间=速度、速度*时间=路程等重新建模。
3	参考变量重要度，有针对性的对客户进行建议。如titanic的家庭成员数量在1~3个时，幸存的可能性更大，那么我们就可以建议客户结伴而行，但同伴不宜太多。

我上面的都学会了，却依然建不好模！

这个原因有很多，预处理做得不完整、算法模型选得不对、参数设置不好，衍生变量没做对，……，这些都是可能的原因，但还有个非常关键的可能因素：

找来的特征变量和目标变量根本就不相关！！！

比如，瓜好不好，和运瓜的车型、摘瓜人叫什么名字是没有什么关系的，如果拿这些特征变量来分析，那是不可能建出好模型滴。这是极端的例子，实践中会有些很多看似相关但其实不相关或相关性很弱的特征变量。所以：

获取真正相关的特征变量数据是建出模型的基础！

可惜，对于这件事，数据挖掘技术无能为力。

如果方向錯了，

停止就是前進

目录 CONTENTS

05

模型应用

模型应用

经过评估，如果模型的预测结果达到预期，则可以将模型应用于实际业务场景中。

右图是使用模型对新数据（没有目标变量Survived）预测titanic幸存者的结果，其结果是乘客幸存即Survived=1的概率。概率越大，表示生还的可能性越高。如第一条记录，预测的结果是11.509%，表示该乘客幸存的概率只有11.509%。

Survived_1_percentage	PassengerId
11.509%	892
34.078%	893
12.999%	894
16.499%	895
55.495%	896
22.879%	897
34.509%	898
28.523%	899
65.75%	900
6.088%	901
2.668%	902
3.984%	903
31.302%	904
2.663%	905
65.241%	906
31.137%	907

模型应用

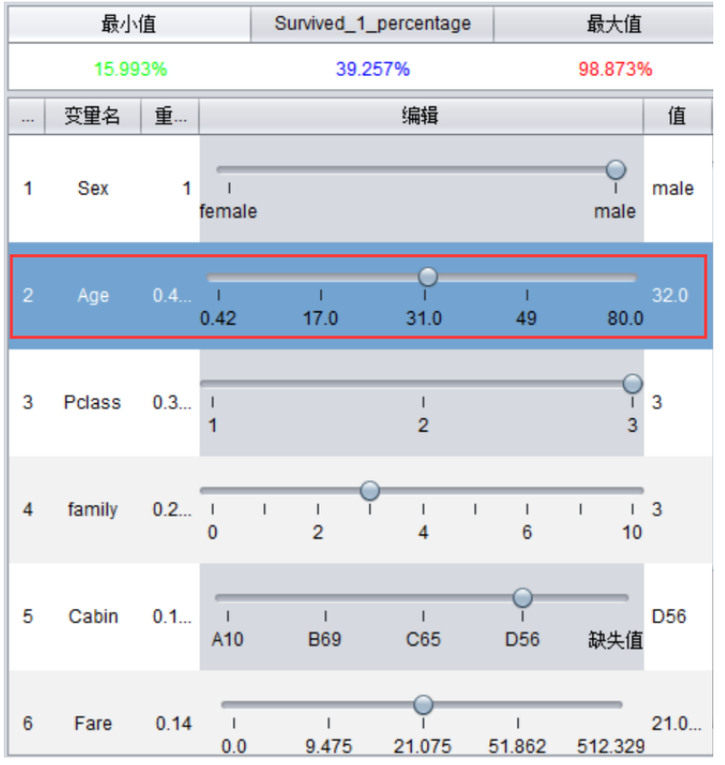
右图是回归模型（房价预测）的预测结果：

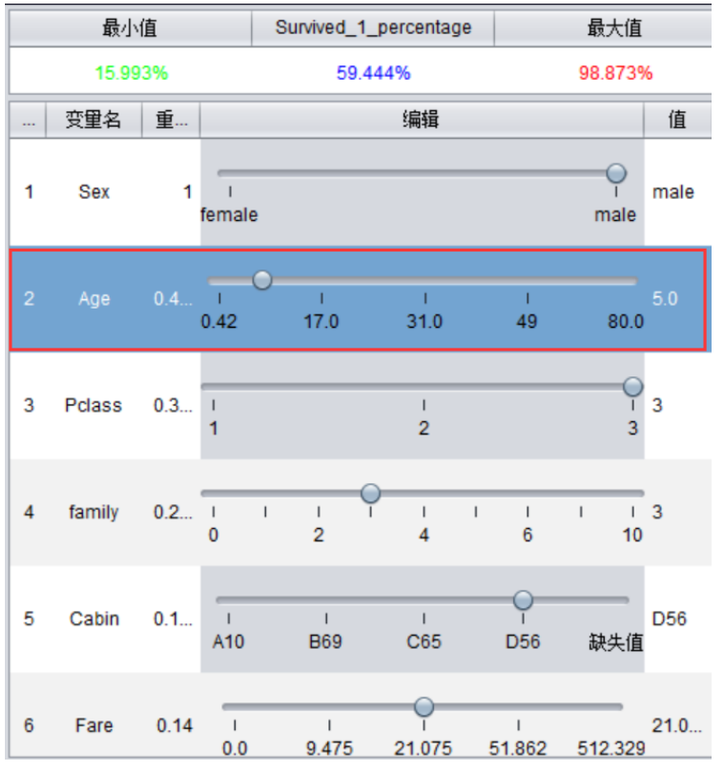
其结果就是预测的房价，如第一条记录，预测的结果是117516.081，表示该房子的预测价格为117516.081美元。

SalePrice_predictvalue	Id	MS
117516.081	1461	
161657.782	1462	
195454.586	1463	
187782.109	1464	
179066.539	1465	
174501.556	1466	
173209.812	1467	
161962.876	1468	
181599.894	1469	
127098.887	1470	
187737.235	1471	
95207.608	1472	
106779.241	1473	
150270.994	1474	
108406.215	1475	
296441.899	1476	

模型应用

还可以对单条数据进行预测，如下图：





Survived_1_percentage	Pclass	Sex	Age	Ticket	Fare	Cabin	Embarked	family
39.257%	3	male	32.0	WE/P 5735	21.075	D56	S	3
59.444%	3	male	5.0	WE/P 5735	21.075	D56	S	3

数据进行预测，年龄为32岁时，幸存的概率只有39.257%，将年龄改为5岁时，幸存的概率则有59.444%。此方法可以应用在营销中，比如降低产品价格可以降低客户流失的可能，则可以据此确定产品的降价幅度。

模型应用

对单个房子数据进行房价预测，如下图：

生活面积为1554时，房价为105441.719，生活面积为5642时，房价为155872.511。用此功能可以为用户提供一些建议。

模型应用

模型还可以在应用程序中，下面是润乾集算器中调用易明智能建模工具生成的模型来预测：

	A	B
1	=file("E:/kaggledata/titanic/titanic_test.csv").import@qtc()	/读入新数据（没有目标变量）
2	=A1.derive(SibSp+Parch:family)	/增加family列
3	E:/kaggledata/titanic/titanic_train.pcf	/指定模型文件
4	=ym_env()	/启动yiming环境
5	=ym_predict(A3)	/根据模型文件生成模型对象
6	=ym_result(A5,A2)	/使用模型预测，生成结果
7	>ym_close(A4)	/关闭yimning环境

A1:

Index	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	892	3	Kelly, Mr. J...	male	34.5	0	0	330911	7.8292	(null)	Q
2	893	3	Wilkes, Mrs...	female	47.0	1	0	363272	7.0	(null)	S
3	894	2	Myles, Mr. T...	male	62.0	0	0	240276	9.6875	(null)	Q

A2:

Index	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	family
1	892	3	Kelly, Mr. J...	male	34.5	0	0	330911	7.8292	(null)	Q	0
2	893	3	Wilkes, Mr...	female	47.0	1	0	363272	7.0	(null)	S	1
3	894	2	Myles, Mr. ...	male	62.0	0	0	240276	9.6875	(null)	Q	0

A6:

Index	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	family	Survived_1...
1	892	3	Kelly, Mr. J...	male	34.5	0	0	330911	7.8292	(null)	Q	0	0.1175322...
2	893	3	Wilkes, Mr...	female	47.0	1	0	363272	7.0	(null)	S	1	0.2679134...
3	894	2	Myles, Mr. ...	male	62.0	0	0	240276	9.6875	(null)	Q	0	0.1441113...

润乾
RAQSOFT

THANKS

创新技术 推动应用进步

www.raqsoft.com.cn