润乾
RAQSOFT

数据挖掘

房价预测

www.raqsoft.com.cn

目录

CONTENTS

01 数据介绍

02 探索和预处理

03 建模与评价

04 模型应用

目录 CONTENTS

01

数据介绍

数据介绍

house_prices房价预测数据集来自于kaggle，是一份用于回归预测的数据集，本节以此数据介绍回归预测的数据挖掘过程（面向对象是小白，大神请绕过）。

house_prices_train.csv：训练集（有目标变量）共有1460条记录，81个变量。
house_prices_test.csv：待测集（无目标变量）共有1459条记录，80个变量。

分析目标：

1.寻找影响房价的因素

2.根据训练集构建模型对待测集数据进行预测

数据介绍

数据的变量数较多，将数据集字典整理成csv文件存储，如下图：

该数据变量比较多，共81个。按类别统计如下：

ID：1

二值变量：4

分类变量：42

计数变量：9

数值变量：20（包含目标变量）

时间日期：5

Variable_Name	Variable_Ty	Variable_Description
Id	ID	
Alley	二值变量	住宅通道的类型
CentralAir	二值变量	是否中央空调
Street	二值变量	街道类型
Utilities	二值变量	公用设施类型
BldgType	分类变量	住宅类型
BsmtCond	分类变量	地下室的一般情况
BsmtExposure	分类变量	地下室采光状况
BsmtFinType1	分类变量	地下室完工区的质量
BsmtFinType2	分类变量	第二个完工区域的质量2
BsmtQual	分类变量	地下室的高度
Condition1	分类变量	靠近主要道路或铁路
Condition2	分类变量	靠近主要道路或铁路2
Electrical	分类变量	电气系统
ExterCond	分类变量	外部材料的状况
Exterior1st	分类变量	房屋外墙
Exterior2nd	分类变量	房屋外墙2
ExterQual	分类变量	外部材料质量
Fence	分类变量	栅栏质量
FireplaceQu	分类变量	壁炉质量
Foundation	分类变量	基础类型

数据介绍

4个二值变量

Alley	二值变量	住宅通道的类型
CentralAir	二值变量	是否中央空调
Street	二值变量	街道类型
Utilities	二值变量	公用设施类型

5个时间日期

GarageYrBlt	日期时间	车库建成年份
MoSold	日期时间	售出月份
YearBuilt	日期时间	原始施工年份
YearRemodAdd	日期时间	改造年份
YrSold	日期时间	售出年份

9个计数变量

BsmtFullBath	计数变量	地下室豪华浴室
BsmtHalfBath	计数变量	地下室浴室
FullBath	计数变量	豪华浴室
HalfBath	计数变量	浴室
BedroomAbvGr	计数变量	卧室数量
Fireplaces	计数变量	壁炉数量
GarageCars	计数变量	车库容量
KitchenAbvGr	计数变量	厨房数量
TotRmsAbvGrd	计数变量	房间总数

20个数值变量

1stFlrSF	数值变量	一楼面积
2ndFlrSF	数值变量	二楼面积
3SsnPorch	数值变量	三季阳台面积
BsmtFinSF1	数值变量	类型1完工区域
BsmtFinSF2	数值变量	类型2完工区域
BsmtUnfSF	数值变量	未完成的地下室面积
EnclosedPorch	数值变量	封闭阳台面积
GarageArea	数值变量	车库的面积
GrLivArea	数值变量	生活区面积
LotArea	数值变量	地块尺寸
LotFrontage	数值变量	相连的街道的英尺数
LowQualFinSF	数值变量	低质量的面积
MasVnrArea	数值变量	砌体饰面区域
MiscVal	数值变量	杂项功能的价值
OpenPorchSF	数值变量	开放式阳台面积
PoolArea	数值变量	泳池面积
SalePrice	数值变量	销售价格，目标变量
ScreenPorch	数值变量	纱窗阳台面积
TotalBsmtSF	数值变量	地下室总面积
WoodDeckSF	数值变量	木甲板面积

剩下的是42个分类变量

这些变量都可能影响房价，按照变量的意义，可以分成这么几类：房子地理位置，房子风格，房子装修，地下室，居住面积，车库，建筑时间等，那我们可以根据自己买房考虑的变量来确定这些变量重要度，也可以把这些作为数据探索和预处的依据。

目录 CONTENTS

02

探索和预处理

数据变量比较多，我们这里不可能每个变量都分析到，因此我们有选择的进行分析，主要介绍数据挖掘的方法，感兴趣的同学可以对变量深入分析和挖掘。

数据探索

1. 根据数据字典检验变量类型

对照整理好的数据字典，修改自动识别的变量类型，如

房价(SalePrice)：计数变量——数值变量

施工年份(YearBuilt)：计数变量——时间日期

壁炉数量(Fireplaces)：分类变量——计数变量

……

变量名	类型
PoolArea	数值变量
SalePrice	数值变量
ScreenPorch	数值变量
TotalBsmtSF	数值变量
HalfBath	计数变量
BedroomAbvGr	计数变量
Fireplaces	计数变量
GarageCars	计数变量
KitchenAbvGr	计数变量
BsmtFullBath	计数变量
BsmtHalfBath	计数变量
FullBath	计数变量
TotRmsAbvGrd	计数变量
GarageYrBlt	时间日期
MoSold	时间日期
YearBuilt	时间日期

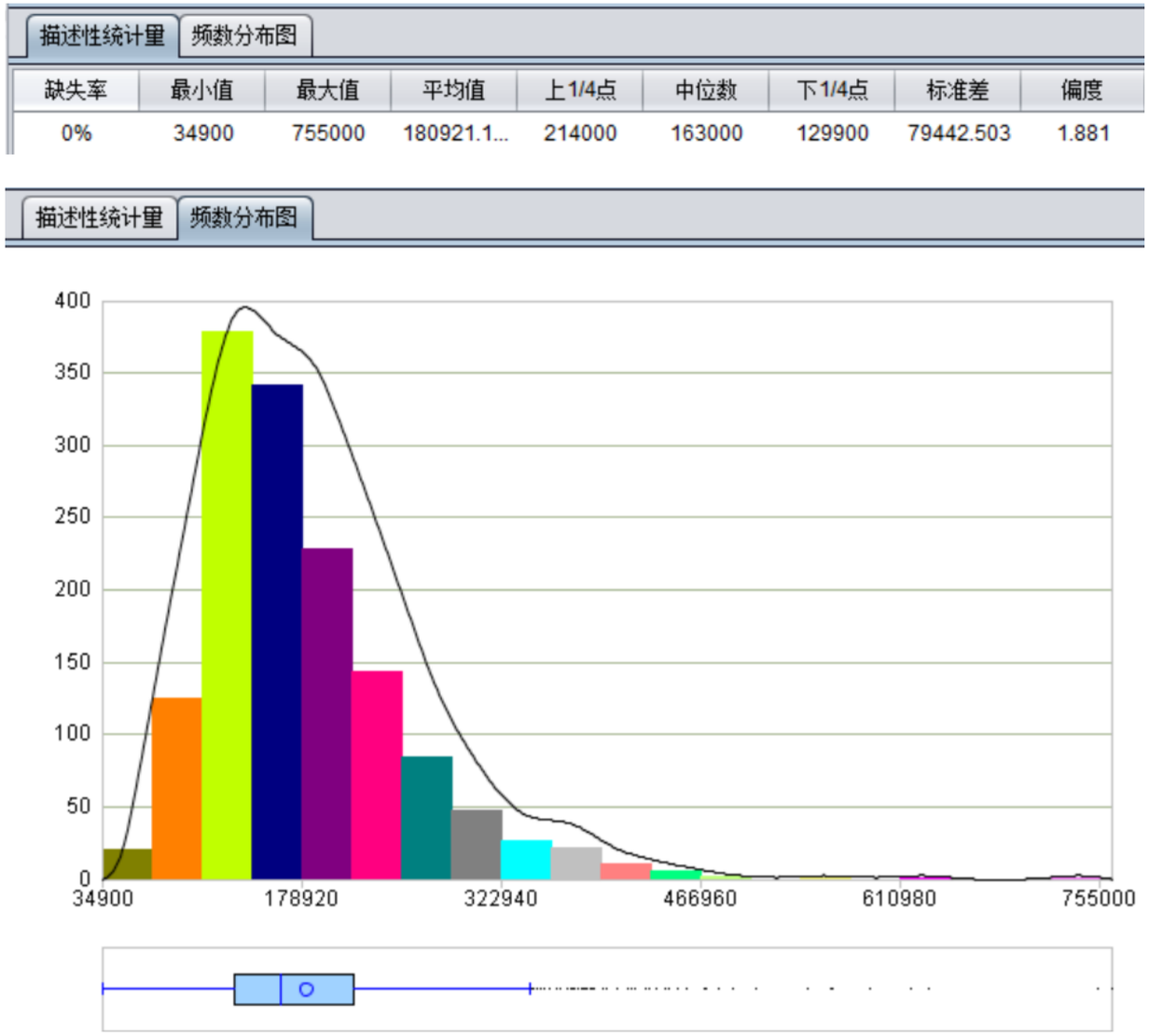
建模工具会错误的把一些数字型的计数变量识别成分类变量，也会把一些数值变量识别成计数变量，还有年份会被识别成计数变量，这些都需要我们进行更改。

2. 目标变量SalePrice

房价，最小值34900，最大值755000，偏度1.881；观察分布图发现，该变量呈右偏，可以进行对数变换使其呈正态分布。

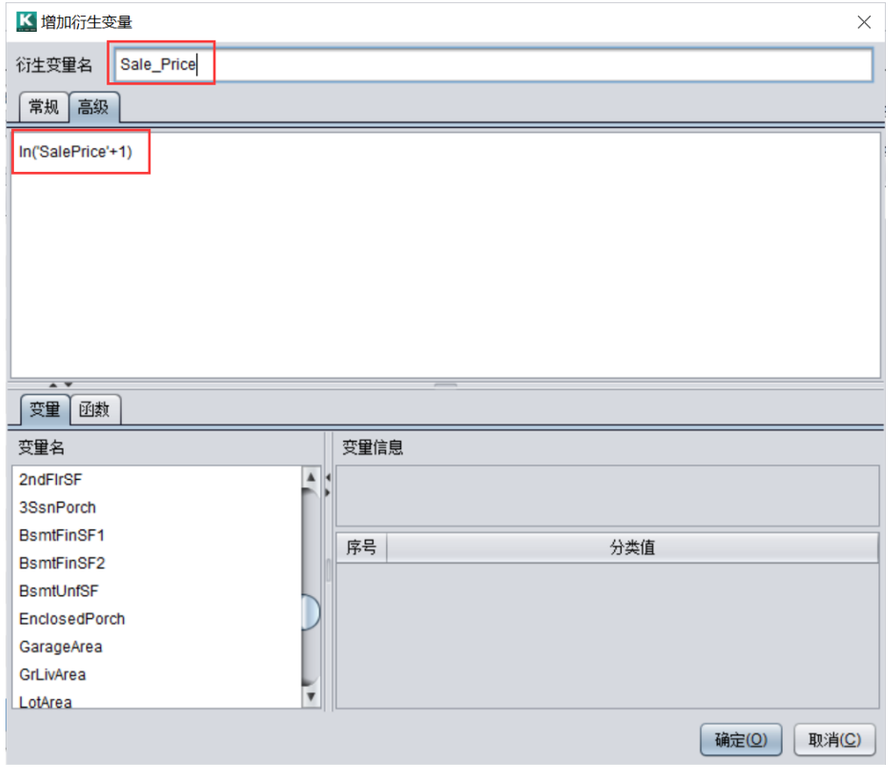
描述性统计量 | 频数分布图

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%	34900	755000	180921.1...	214000	163000	129900	79442.503	1.881



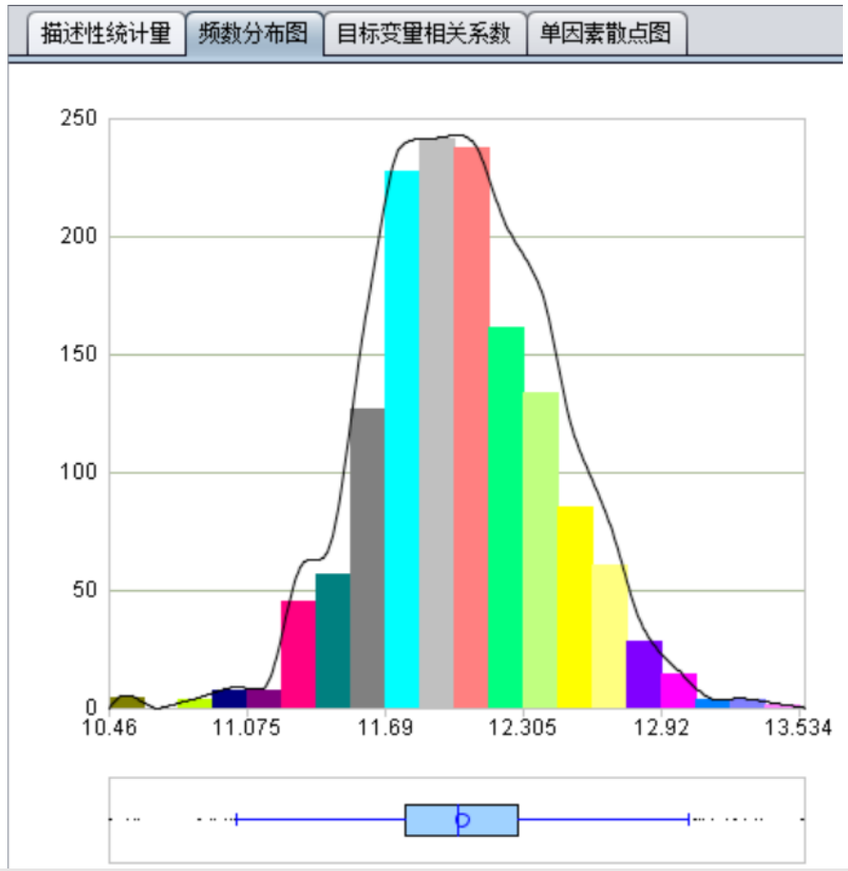
数据预处理——数据纠偏

使用对数变换对SalePrice进行纠偏



描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

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%	10.46	13.534	12.024	12.274	12.002	11.775	0.399	0.121

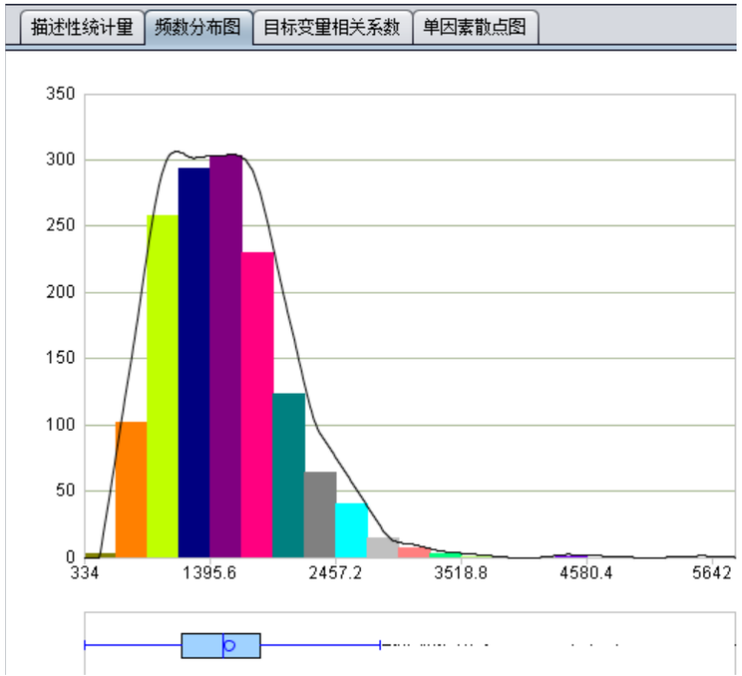


观察变换后的变量，基本呈正态分布了，有利于模型建模。

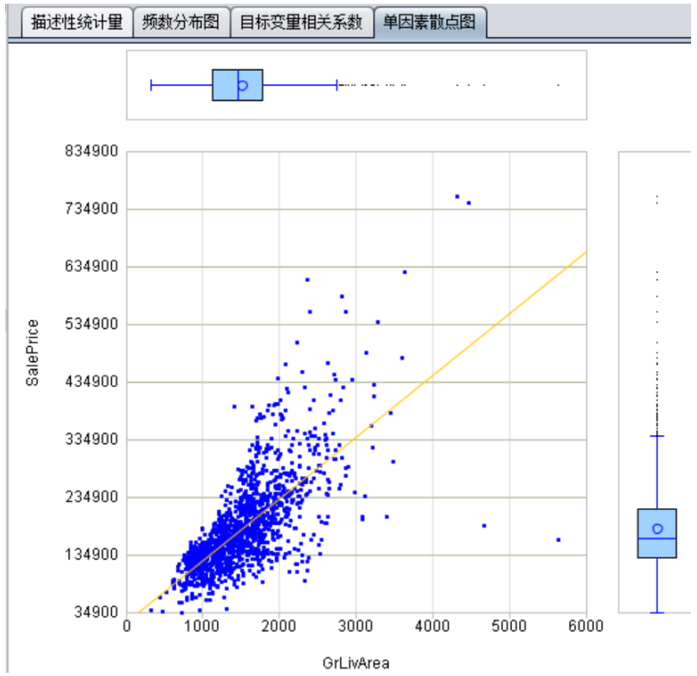
注意：智能建模工具会自动对数值变量做纠偏和规范化等处理工作，不需要手动完成，只要把变量类型设置正确即可。这里只是为了提醒读者数据需要使用对数变换进行纠偏处理。

数据探索-分布分析、统计量分析、相关性分析

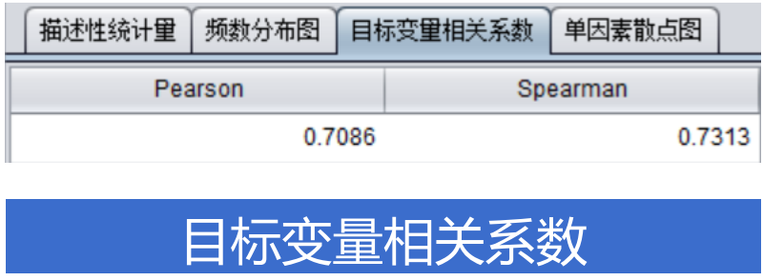
与目标变量分布类似的特征变量，也可以采用相同的方法，如生活面积GrLivArea



频数分布图



单因素散点图



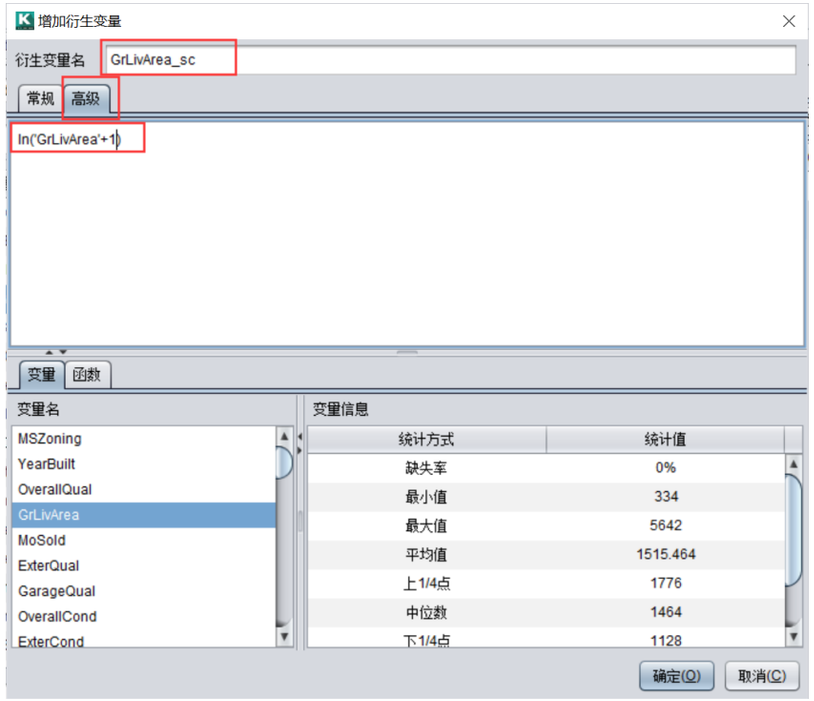
目标变量相关系数

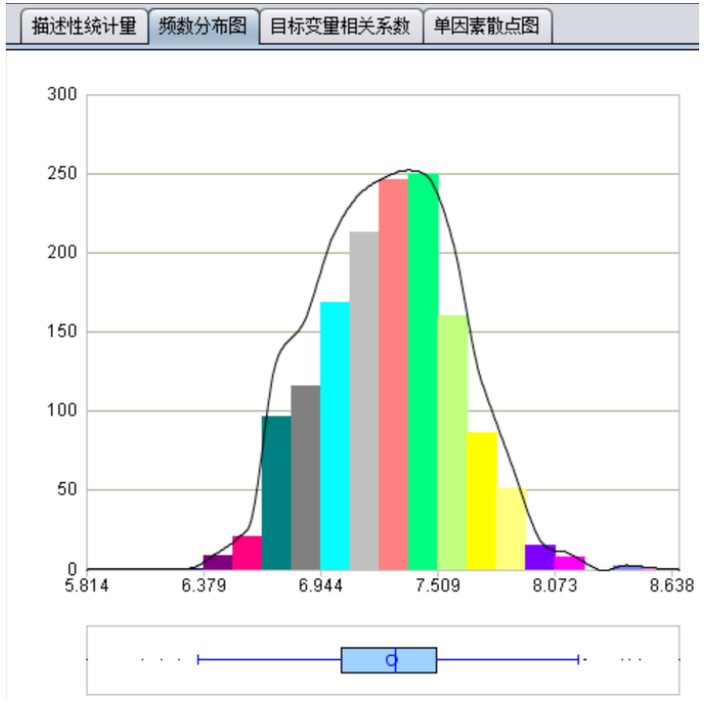
观察发现，我们不仅可以看到数据的分布情况，还可以看到它与目标变量的相关系数和单因素散点图，两个相关系数都大于0.7说明该变量与房价相关性很强，这一点在散点图上同样可以看出，生活面积越大，房价呈上升趋势，因此该变量的重要性可见一斑。

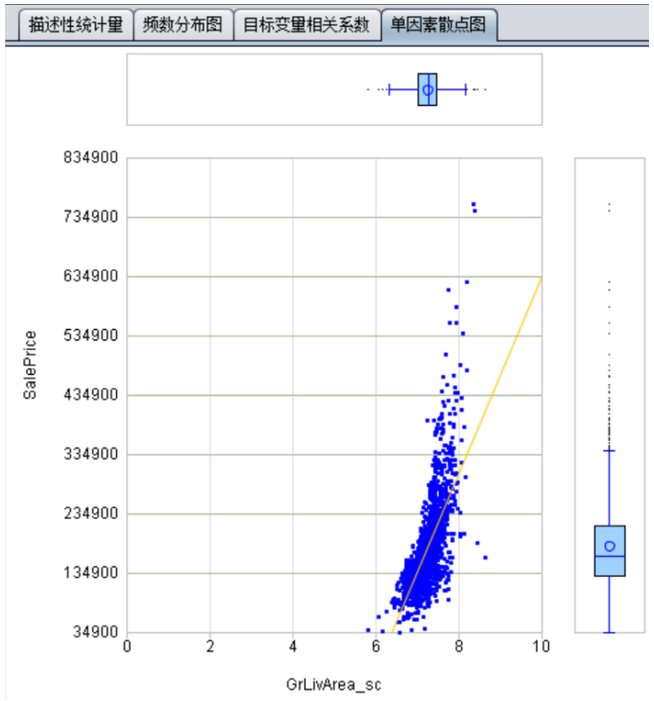
数据预处理 - 数据纠偏

使用对数变换对生活面积GrLivArea进行纠偏

这个数据中具有和GrLivArea分布类似的变量还有地下室面积TotalBsmtSF、一楼面积1stFlrSF、相连的街道的英尺数LotFrontage 等等，都可以使用同样的方法处理。







描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

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%	5.814	8.638	7.269	7.483	7.29	7.029	0.333	-0.006

描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

Pearson	Spearman
0.6951	0.7313

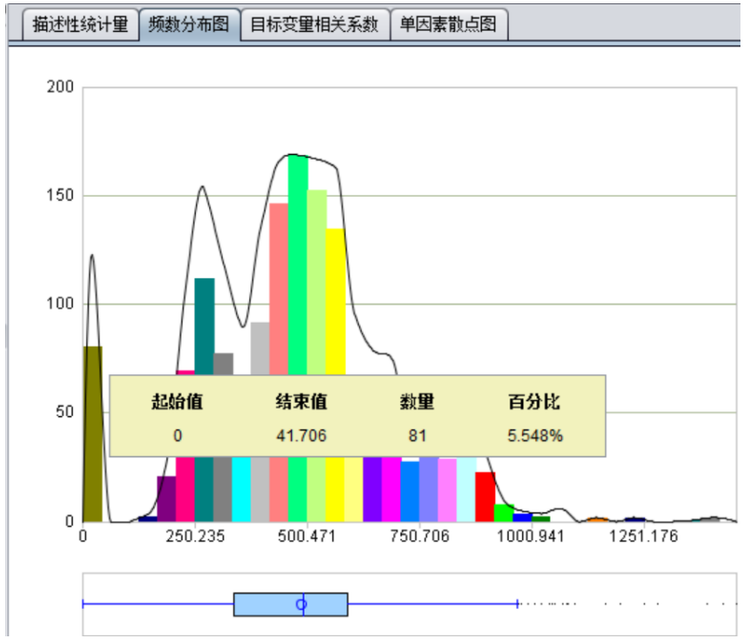
数据纠偏后，偏度变为-0.006，几乎不偏斜，相关系数和单因素散点图同样非常高，说明该变量相较于原变量对建模更有利。

数据探索 - 分布分析、统计量分析、相关性分析

车库面积GarageArea

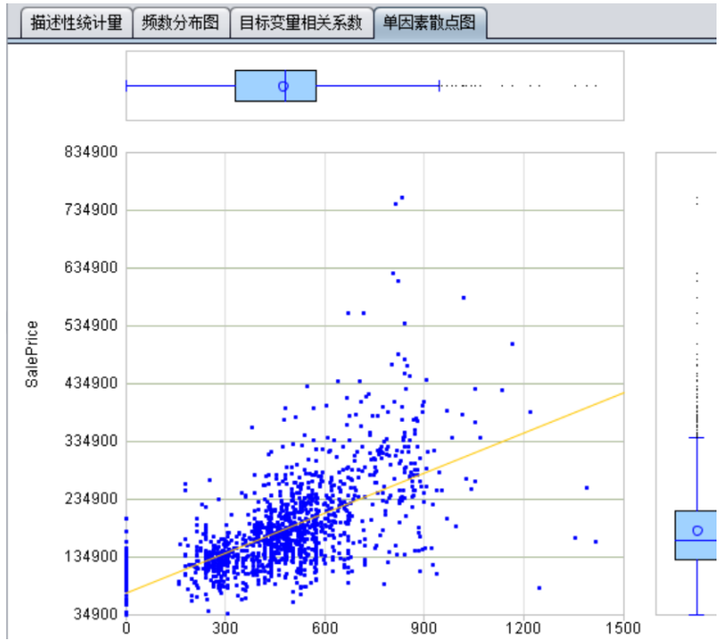
描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

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	1418	472.98	576	480	330	213.8...	0.18



描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

Pearson	Spearman
0.6234	0.6494



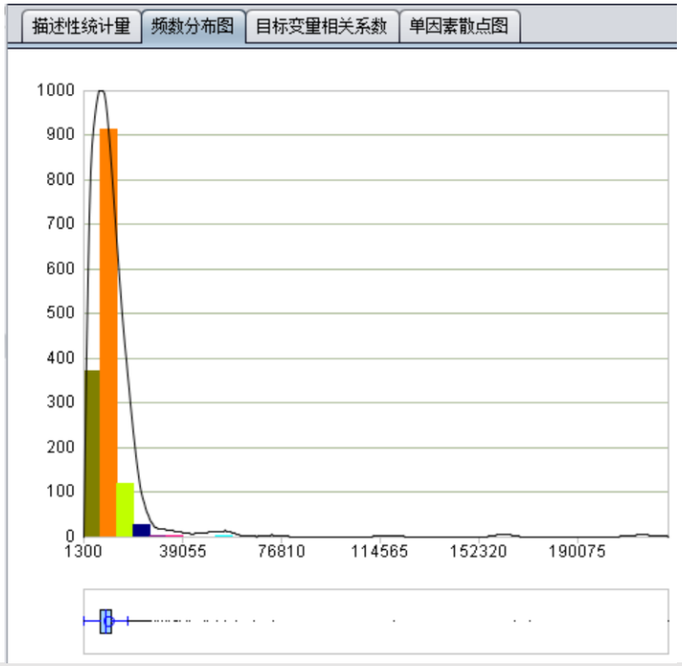
观察车库面积，发现有81个房子的车库面积是0，也就是说这些房子没有车库，整体分布类似正态分布而且看相关系数和单因素散点图说明该变量与房价呈正相关，不需要处理。

数据探索 - 分布分析、统计量分析、相关性分析

地块面积LotArea

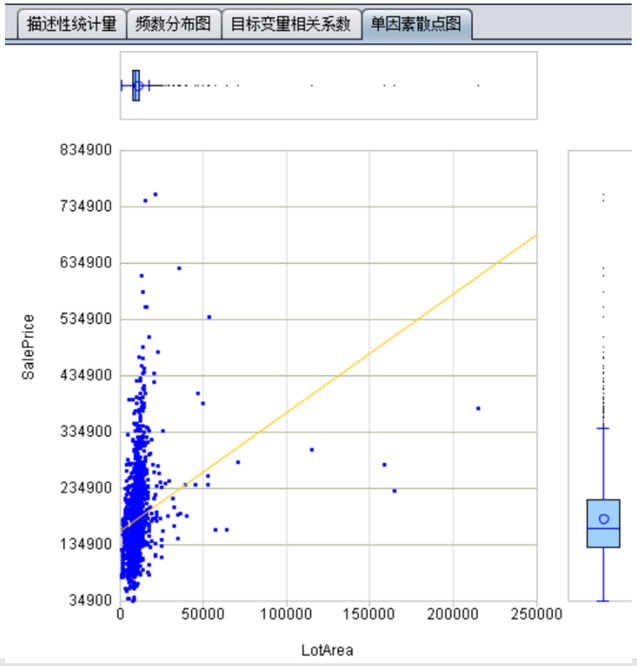
描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

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%	1300	215245	10516.828	11600	9477	7540	9981...	12.195



描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

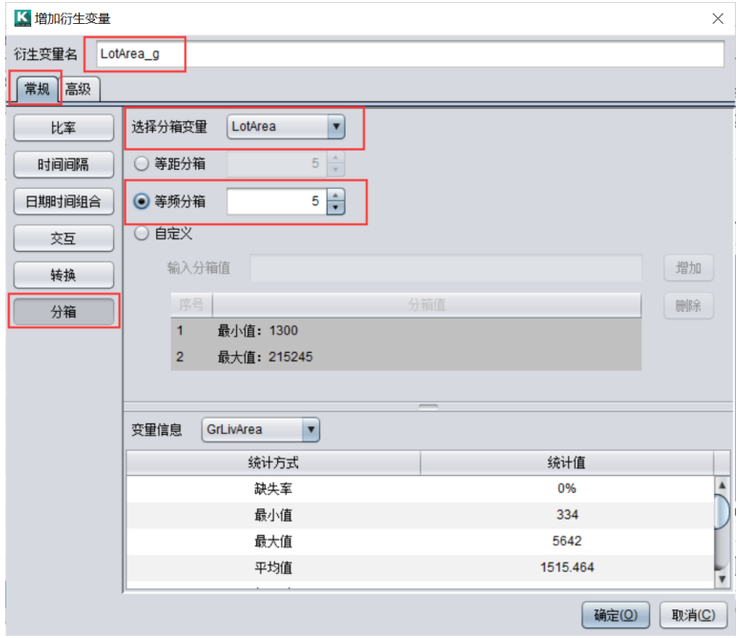
Pearson	Spearman
0.2638	0.4565

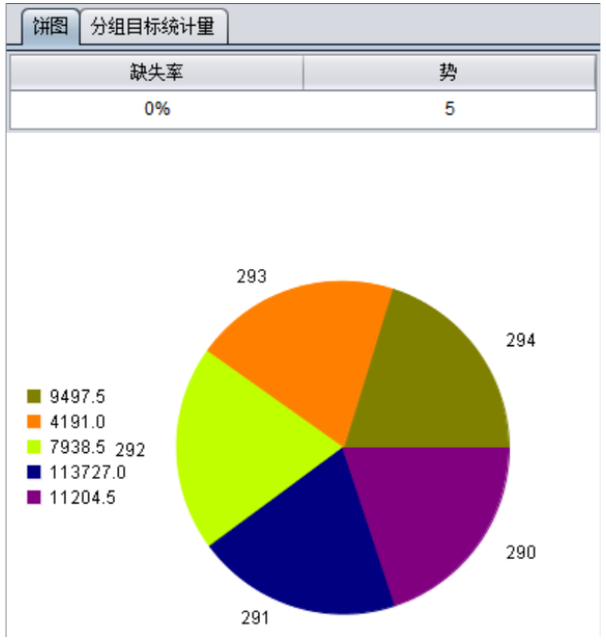


观察发现，LotArea的偏度为12.195，太大了，很难纠偏了，因此我们可以考虑使用等频分箱的方法来将其离散化为分类变量，再观察两个相关系数，Pearson系数只有0.2638，Spearman相关系数则有0.4565，说明该变量与房价不是线性相关，但依旧呈现出单调的趋势即地块面积越大，房价越高，但非线性的关系。

数据预处理 - 连续数据离散化

使用等频分箱的方法对地块面积LotArea进行离散化





饼图 | 分组目标统计量

分类变量	频率	平均值	标准差	中位数	最小值	最大值
4191.0	293	139894...	49098....	143000	37900	392500
7938.5	292	152285...	47519....	194700	34900	392000
9497.5	294	173624...	61650....	161500	35311	424870
11204.5	290	196385...	69644....	222500	64500	466500
113727.0	291	242925...	108200...	755000	81000	755000

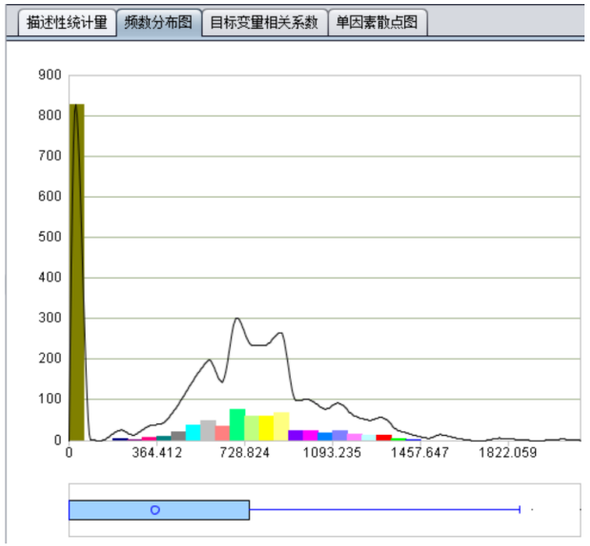
等频分箱后，观察分组后的统计量，随着地块面积的增大，房价的平均值明显的上升，这个变量应该是有效的。

数据探索-分布分析、统计量分析、相关性分析

二楼面积2ndFlrSF

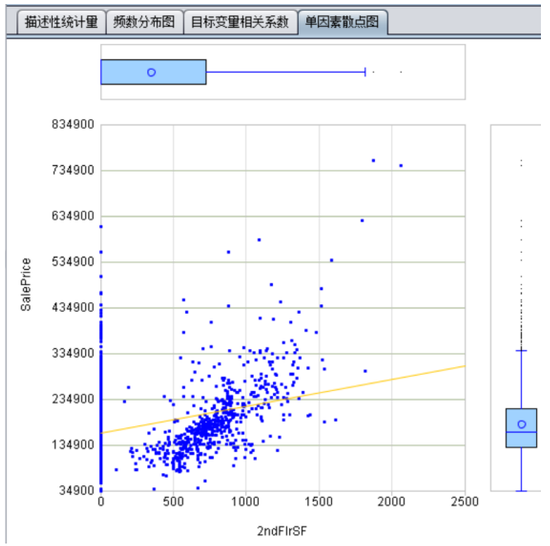
描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

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	2065	346.992	728	0	0	436.5...	0.812



描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

Pearson	Spearman
0.3193	0.3569

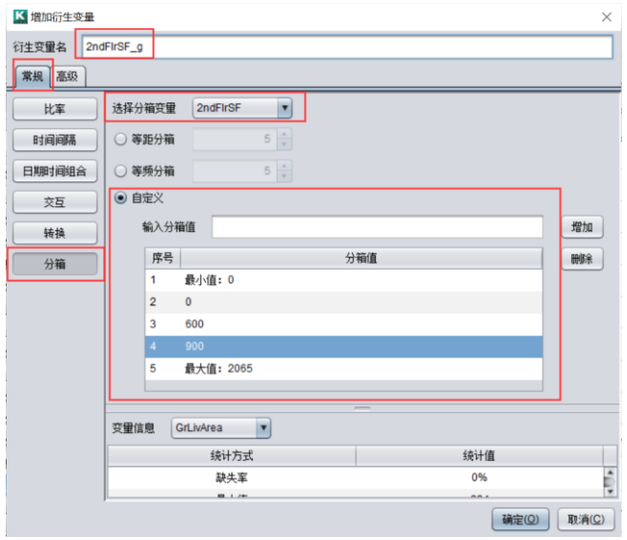


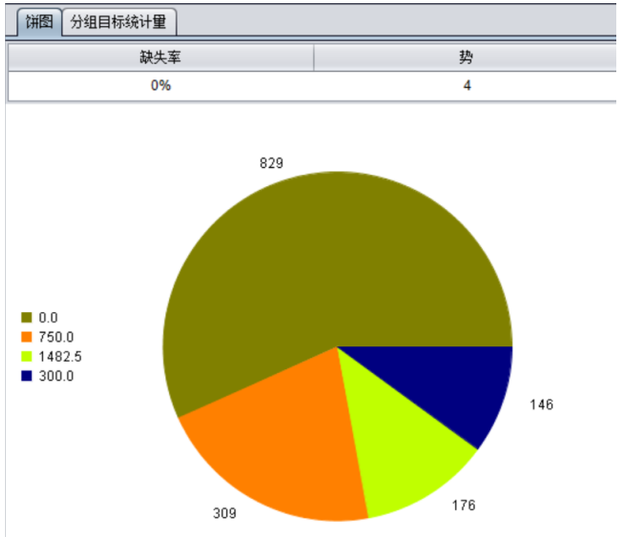
观察发现， 2ndFlrSF的偏度为0.812，偏度不算大，但观察分布图后可以看到有相当一部分的房子二楼面积是0，也就是平房，这时候我们就要考虑分箱，区分平房和2层楼，而且从散点图可以看到，二层楼的房价是随着二层的面积增加呈上升趋势的，因此把他分成四部分比较合理，即0,0~600,601~900,900以上。

数据预处理 - 连续数据离散化

和2ndFlrSF分布类似的变量还有木地板面积WoodDeckSF等等，我们可以根据不同的数据分布情况进行分箱（怎么分，分多少）。

使用自定义分箱的方法对二楼面积2ndFlrSF进行离散化





饼图 | 分组目标统计量

分类变量	频率	平均值	标准差	中位数	最小值	最大值
0.0	829	171383.671	73921.072	194500	34900	611657
300.0	146	133242.007	58199.028	141500	37900	451950
750.0	309	184877.495	53385.366	130500	40000	556581
1482.5	176	258451.159	102663.157	265979	90000	755000

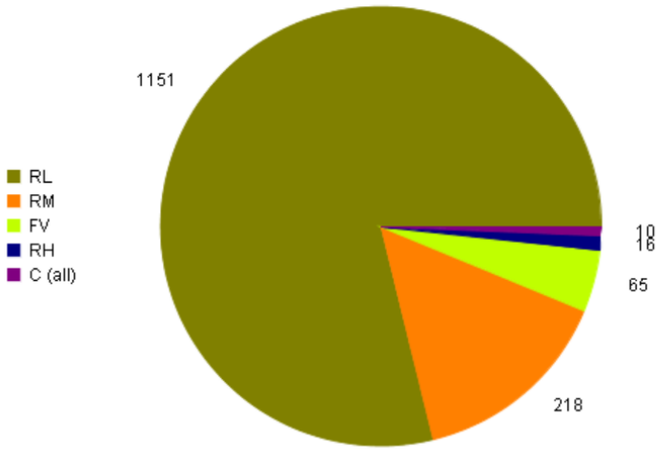
自定义分箱后，观察分组后的统计量，平房的房价应该更多考虑其他的方面，而二层楼的则很好的体现除了面积越大房价越高，提取了其中的规律。

数据探索-分布分析

城市分区MSZoning

饼图 | 分组目标统计量

缺失率	势
0%	5



饼图 | 分组目标统计量

分类变量	频率	平均值	标准差	中位数	最小值	最大值
C (all)	10	74528.0	33791.092	108000	34900	133900
FV	65	214014.0...	52369.662	205950	144152	370878
RH	16	131558.3...	35714.118	140000	76000	200000
RL	1151	191004.9...	80766.341	180000	39300	755000
RM	218	126316.83	48521.69	93000	37900	475000

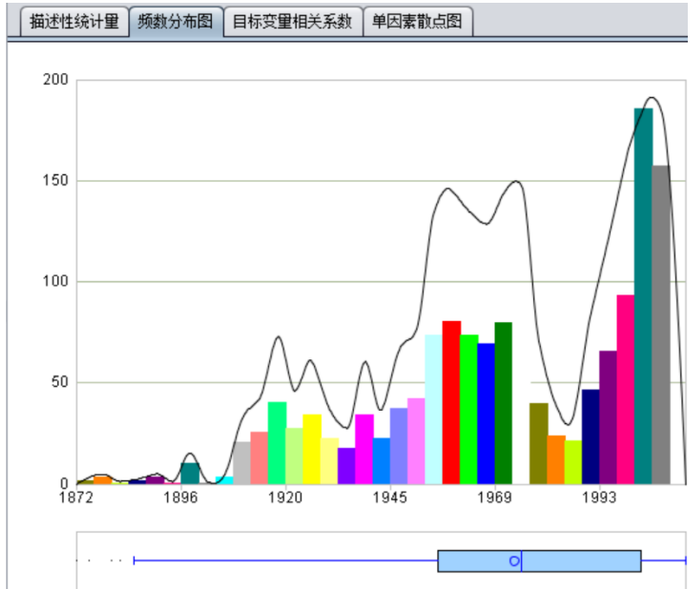
我们买房子首先考虑的就是位置，地段。观察MSZoning，RL的房子最多，其次是RM,RH,FV和C(all)，观察分组后的房价均值，可以发现各个分区的价格差别还是很明显的，因此和我们的直觉一样，位置很关键。对于这样的分类变量，通常采用one-hot编码进行预处理，建模工具预处理时会进行这一步，因此不需要我们来手动完成。

数据探索-分布分析、统计量分析、相关性分析

施工日期YearBuilt

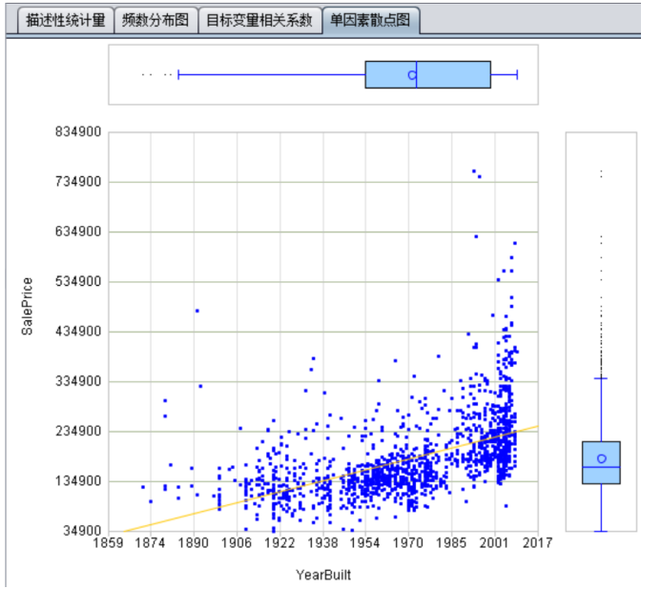
描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

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%	1872	2010	1971	2000	1973	1954	-	-



描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

Pearson	Spearman
0.5229	0.6528



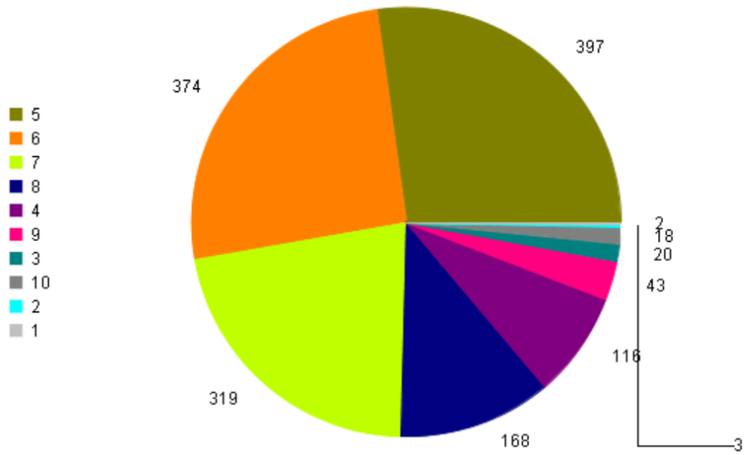
买房子当然要考虑房子的时间，旧房子通常不如新房子贵，可以看到最早的房子是1872年的(历史悠久啊)，但是无论是相关系数和散点图都显示出了，年代和房价基本正相关这样一个关系。

数据探索-分布分析

整体质量情况OverallQual

饼图 | 分组目标统计量

缺失率	势
0%	10



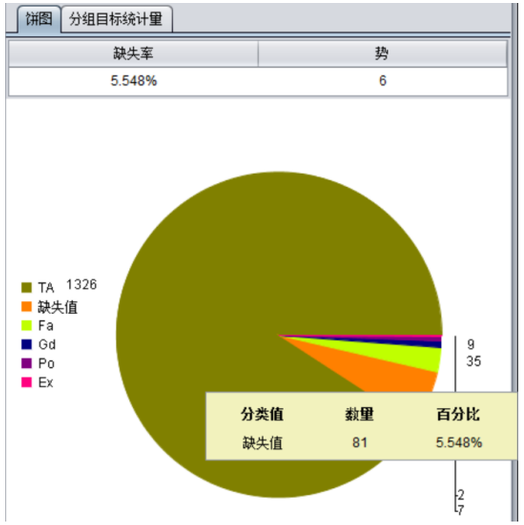
饼图 | 分组目标统计量

分类变量	频率	平均值	标准差	中位数	最小值	最大值
1	2	50150.0	15344.217	61000	39300	61000
2	3	51770.333	14254.201	60000	35311	60000
3	20	87473.75	24688.567	37900	37900	139600
4	116	108420.655	29022.004	96500	34900	256000
5	397	133523.348	27107.331	114500	55993	228950
6	374	161603.035	36090.183	194700	76000	277000
7	319	207716.423	44466.259	194500	82500	383970
8	168	274735.536	63898.902	254000	122000	538000
9	43	367513.023	81278.175	313000	239000	611657
10	18	438588.389	159785.491	451950	160000	755000

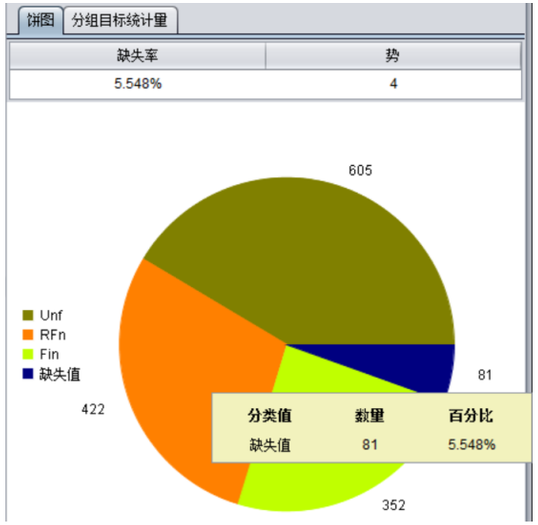
这个变量凭直觉就很重要的，因为它给房子评了10个等级，等级越高，质量越好，房价当然越高，查看分组目标统计量也很容发现这个关系。

数据探索-分布分析、缺失值分析

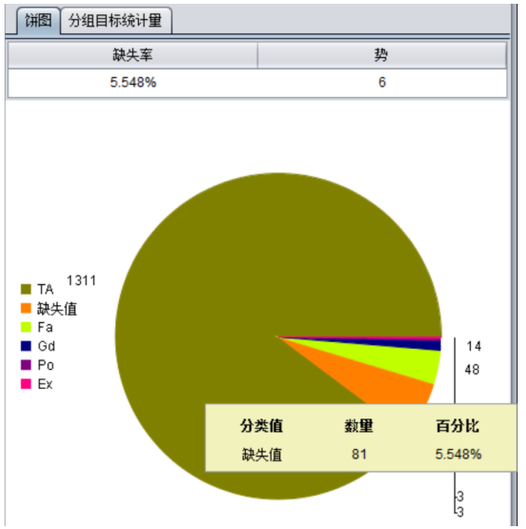
车库系列GarageX



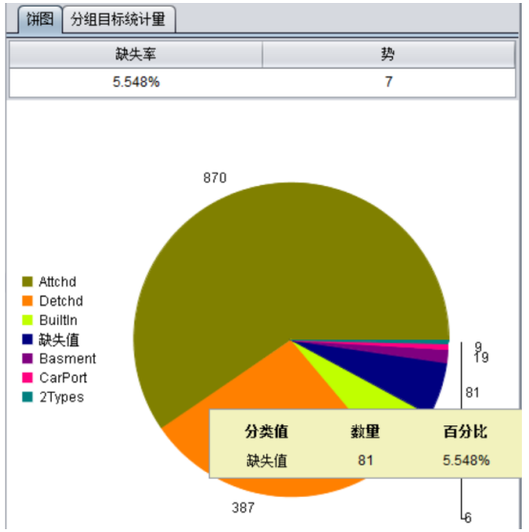
车库状况GarageCond



车库装饰GarageFinish



车库质量GarageQual



车库位置GarageType

车库年份GarageYrBlt

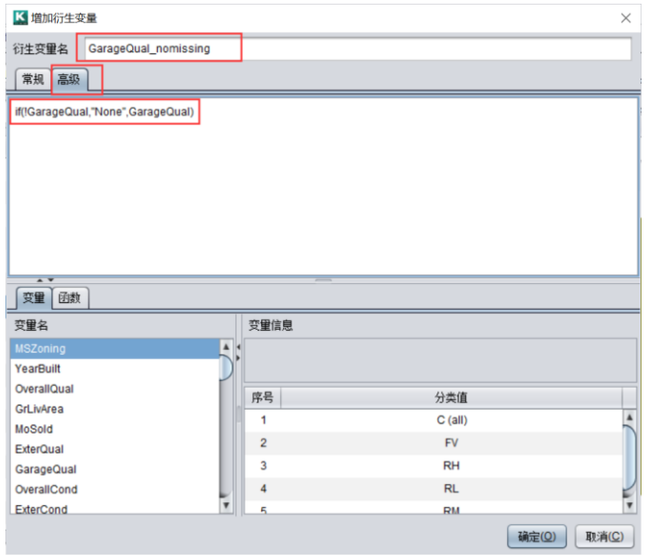
描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

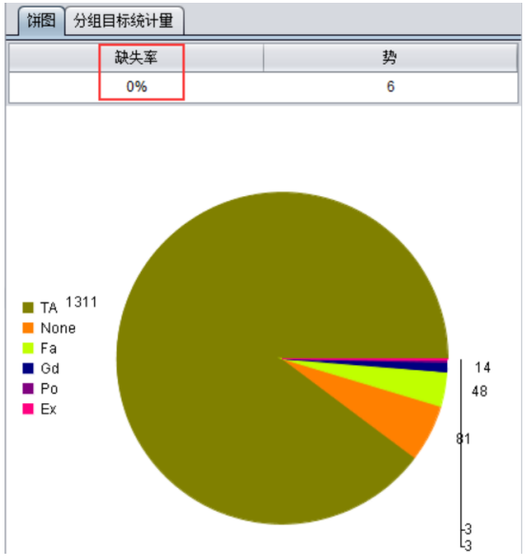
缺失率	最小值	最大值	平均值	上1/4点	中位数	下1/4点	标准差	偏度
5.548%	1900	2010	1978	2002	1980	1961	-	-

前边分析车库面积时，发现有81个房子是没有车库的，所以关于车库的一些属性就缺失了，因此针对这些变量，分类变量的缺失值设为None，成为单独1类，而年份的话我们把它们都设置成车库建成的最早值即1900。

数据预处理 - 缺失值处理

以车库质量GarageQual为例增加衍生变量GarageQual_nomissing，填补缺失值





饼图 | 分组目标统计量

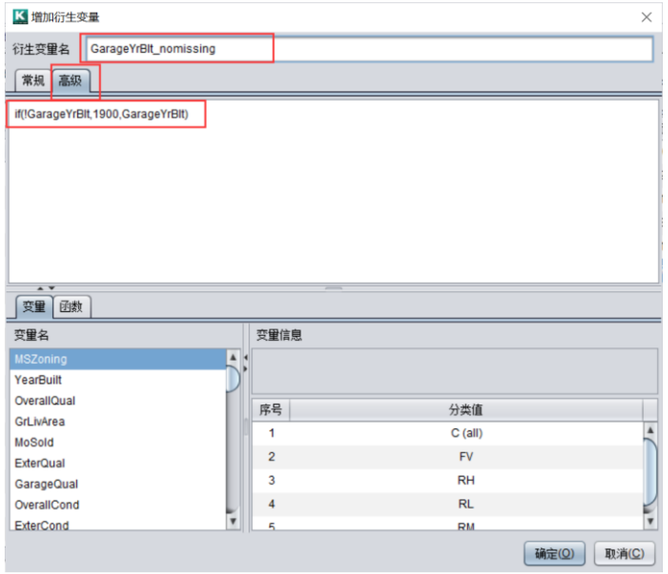
分类变量	频率	平均值	标准差	中位数	最小值	最大值
Ex	3	241000.0	202680.167	475000	120500	475000
Fa	48	123573.354	42971.441	109500	64500	256000
Gd	14	215860.714	74126.739	157500	90350	325000
None	81	103317.284	32815.023	60000	34900	200500
Po	3	100166.667	35143.752	137000	67000	137000
TA	1311	187489.836	78774.949	131400	35311	755000

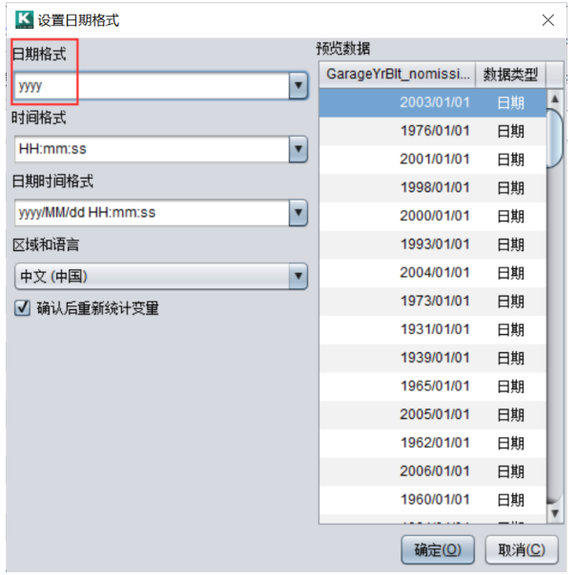
这份数据的车库系列的缺失率不是特别少，智能建模工具内部也会自动进行该处理，这里只是提醒大家填补缺失值的时候需要根据数据特点来填补。

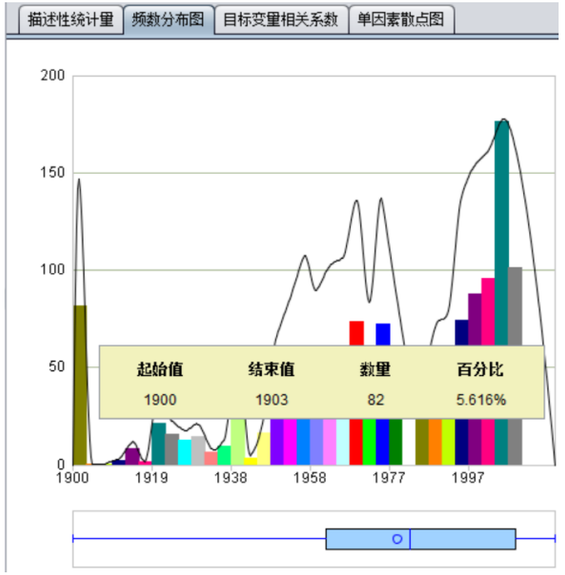
数据预处理 - 缺失值处理

和车库系列类似的变量还有泳池PoolX系列、壁炉FirePlaceX系列、地下室BsmtX系列，使用相同的方法填补就可以了。

车库年份GarageYrBlt增加衍生变量GarageYrBlt _nomissing，使用1900填补缺失值







对于时间日期变量的缺失值，智能建模有单独的缺失值填补方式，这里就填补成1900就可以了。

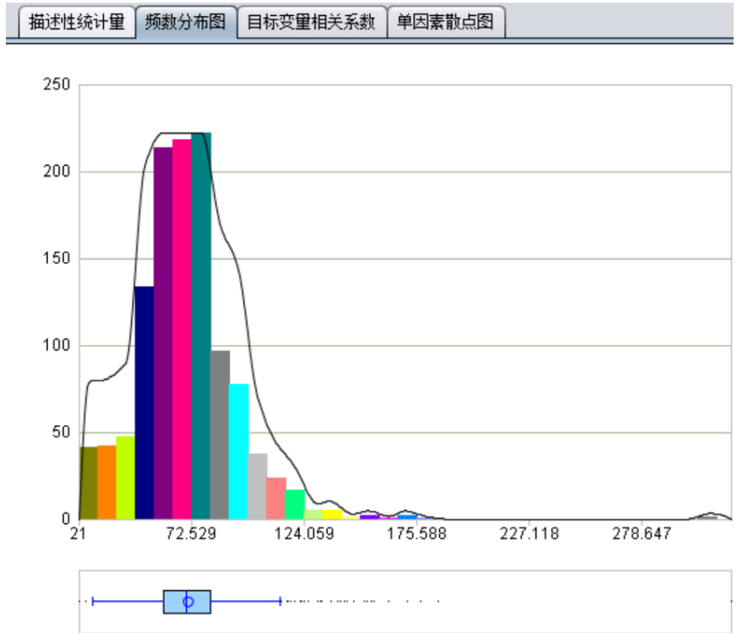
注意：填补后我们注意到新增的衍生变量成了分类变量，我们需要将其改为时间日期类型。填补后我们发现1900年的车库变多了。缺失值没有了。

数据探索-分布分析、缺失值分析

街道长度LotFrontage

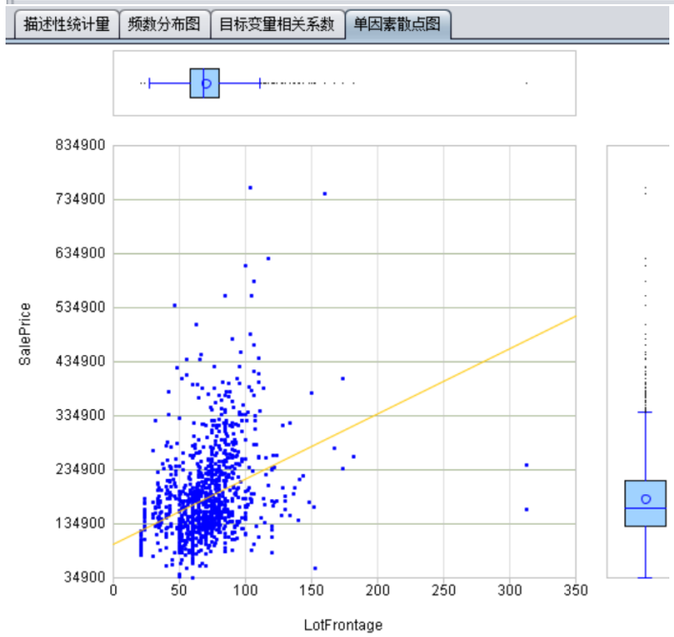
描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

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
17.74%	21	313	70.05	80	69	59	24.285	2.161



描述性统计量 | 频数分布图 | 目标变量相关系数 | 单因素散点图

Pearson	Spearman
0.3518	0.4098



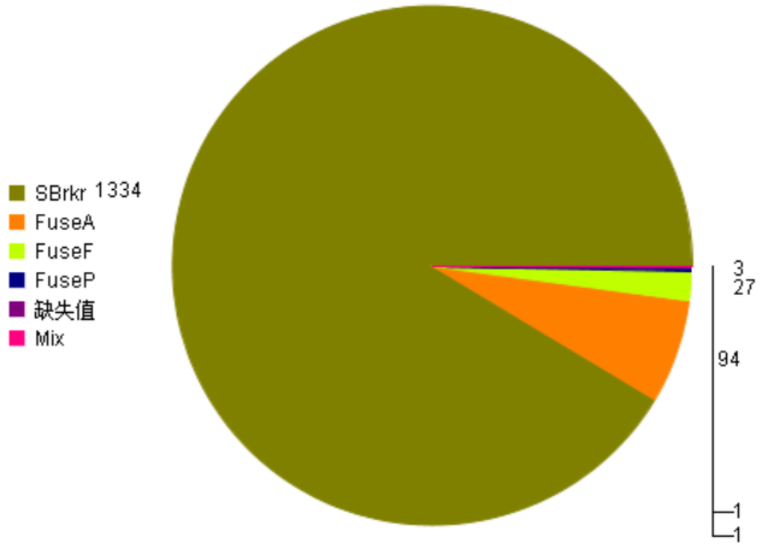
这种分布的缺失值的填补是比较常见的，由于其偏度比较大，填补的时候使用中位数填补即可。智能建模工具对于这种缺失值会自动进行处理，因此这里我们就不再处理了。

数据探索-分布分析、缺失值分析

电器系统Electrical

饼图 | 分组目标统计量

缺失率	势
0.068%	6



饼图 | 分组目标统计量

分类变量	频率	平均值	标准差	中位数	最小值	最大值
缺失值	1	167500.0	-	167500	167500	167500
FuseA	94	122196.894	37511.377	79500	34900	239000
FuseF	27	107675.444	30636.507	128000	39300	169500
FuseP	3	97333.333	34645.827	82000	73000	137000
Mix	1	67000.0	-	67000	67000	67000
SBrkr	1334	186825.113	79856.458	142000	37900	755000

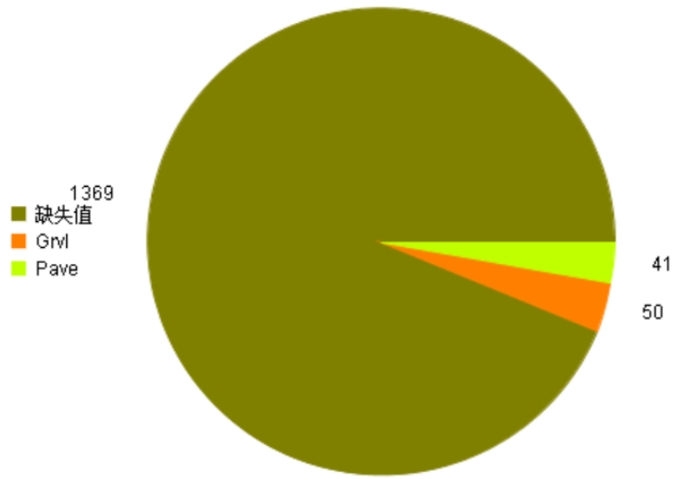
这种分类变量缺失值很少的情况，我们可以选择删除这一条记录，或者使用众数来填补，智能建模工具通常是使用众数填补。

数据探索-缺失值分析

通道类型Alley

饼图 | 分组目标统计量

缺失率	势
93.767%	3



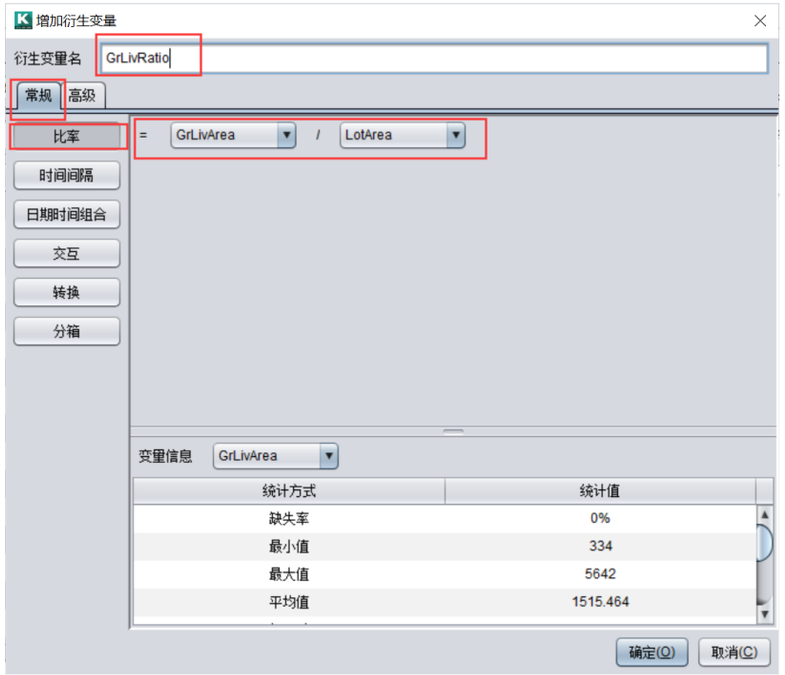
4	Alley	二值变里	

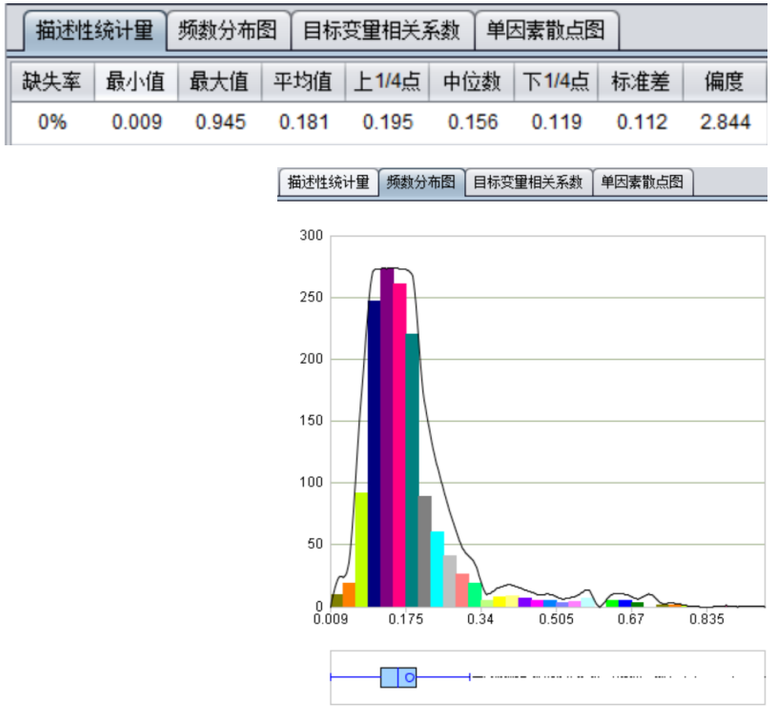
这种缺失数量特别多的变量，通常用以下两种方式处理，第一种就是直接舍弃；第二种是将缺失值视为另一个类别，将其设为None。因为我们买房的时候也很少考虑通道类型是柏油的或者沥青等等的，所以将其删除即可。

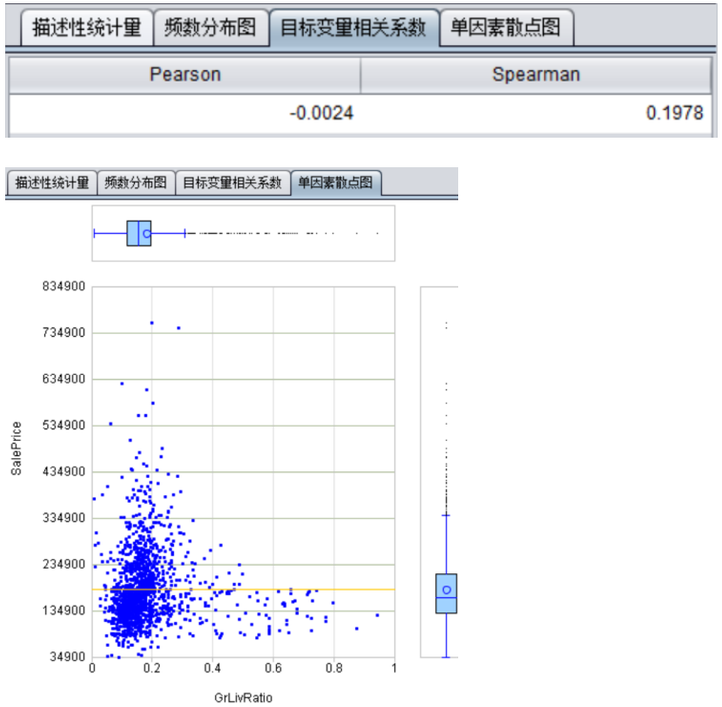
这里介绍的缺失值填补都是简单的填补方式，通常也是比较有效的方法。但智能建模工具有智能填补的方式，它是通过建立模型，对缺失值进行填补，填补更加有效和有针对性。

数据预处理-增加衍生变量、变量交互

用生活区面积GrLivArea/地块尺寸LotArea生成生活区比例



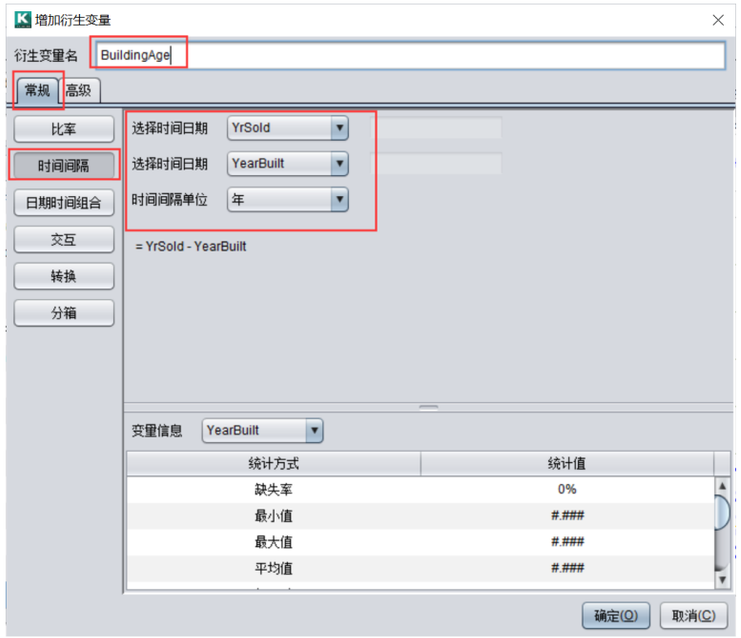


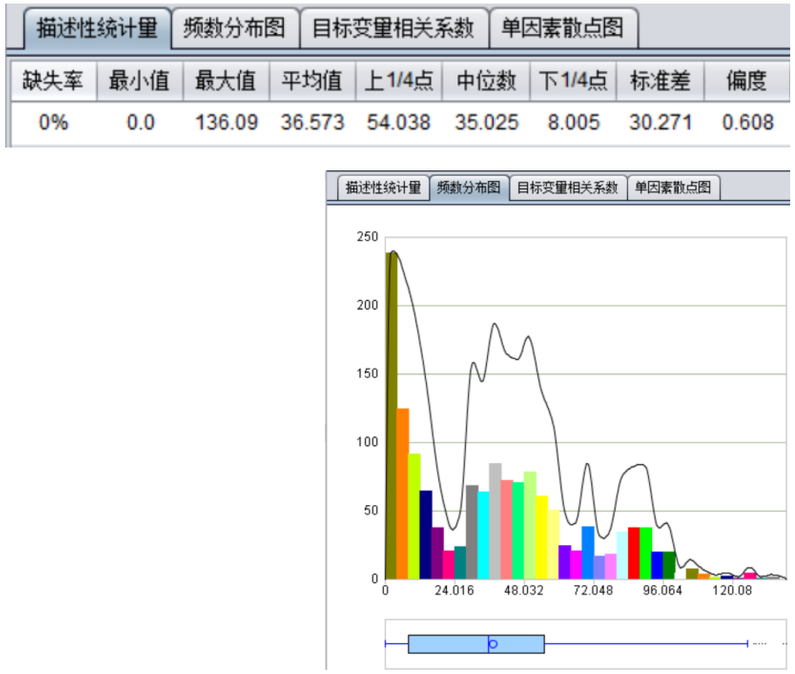


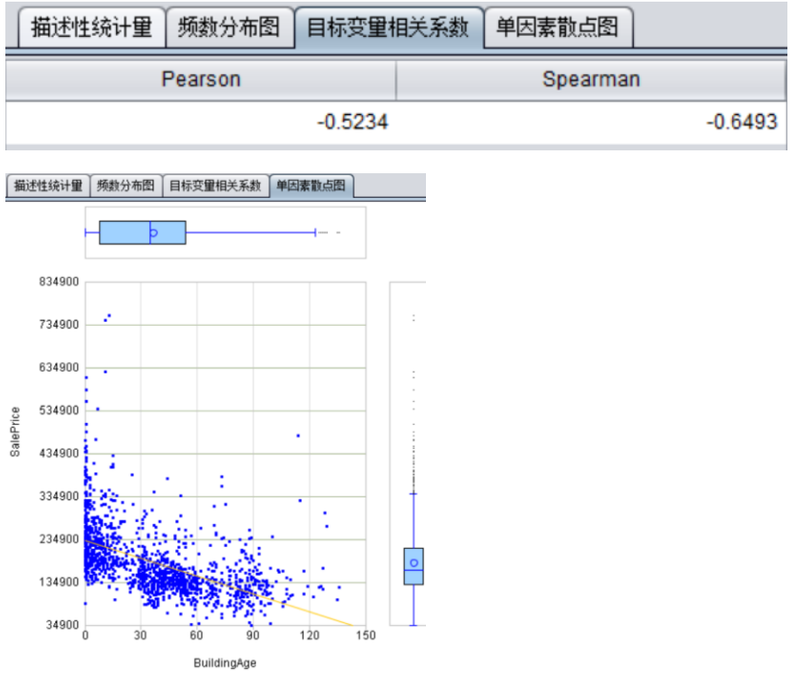
观察这个衍生变量，发现它有偏严重，应该进行纠偏，而且相关系数很小，散点图同样显示相关度很低，初步推断这个变量用处不大，这里主要是介绍两个变量之间可以使用比率方式进行交互生成衍生变量，这是智能建模工具内部暂时所不具备的功能。

数据预处理 - 增加衍生变量、计算时间间隔

计算售出年份YrSold与施工年份YearBuilt的时间间隔最为建筑年龄BuildingAge







观察这个衍生变量，最小值0，最大值136.09，相关系数小于0，但绝对值都大于0.5，说明它与房价负相关，看散点图也发现相同的规律，因此该变量应该是比较重要的，但和YearBuilt 有相同的作用，这里也只是介绍这个方法，真的预处理的时候不需要这步处理。

数据预处理-去除冗余变量

经过预处理后，我们去除以下冗余变量：

1.地块面积LotArea：生成衍生变量LotArea_g了，不需要。

2.二楼面积2ndFlrSF：生成衍生变量2ndFlrSF_g了，不需要。

3.车库年份GarageYrBlt：生成衍生变量GarageYrBlt_nomissing了，不需要。

4.通道类型Alley：缺失率过高且与房价关系不大，舍弃

如果使用以上相同的方法生成了衍生变量，则原变量可以去除；
如果我们不确定生成的衍生变量是否比原变量更有利于建模且时间允许的情况下，则保留原变量，一起交给建模工具来处理并建模也是可以的。

数据探索和预处理汇总表

序号	变量名	探索内容	探索结果	预处理内容	预处理结果
1	SalePrice	分布分析、统计量分析	分布偏斜，需纠偏	对数纠偏	不再偏斜
2	GrLivArea	分布分析、统计量分析、相关性分析	分布偏斜，与SalePrice正相关	对数纠偏	不再偏斜
3	GarageArea	分布分析、统计量分析、相关性分析	有些房子无车库		
4	LotArea	分布分析、统计量分析、相关性分析	偏斜严重	等频离散化	生成分量变量
5	2ndFlrSF	分布分析、统计量分析、相关性分析	有些房子无二楼	自定义离散化	生成分量变量
6	MSZoning	分布分析，分组统计	分区与房价关系紧密		
7	YearBuilt	分布分析、统计量分析、相关性分析	与SalePrice正相关		
8	OverallQual	分布分析，分组统计	房子整体质量与房价关系紧密		
9	GarageX	分布分析，缺失值分析	缺失是因为无车库	缺失值作为新的类别	衍生无缺失值变量
10	GarageYrBlt	分布分析，缺失值分析	缺失是因为无车库	最早年份填补	衍生无缺失值变量
11	LotFrontage	分布分析，缺失值分析	分布偏斜，需填补缺失值		
12	Alley	分布分析，缺失值分析	缺失值过多且不重要，舍弃		
13	GrLivArea/LotArea			变量交互	衍生比率变量
14	YrSold-YearBuilt			变量交互	衍生时间间隔

注意：上述变量的预处理，只有黄色和蓝色部分是需要我们手动进行的，其中的变量交互主要是介绍数据预处理方法，生成的两个衍生变量对房价预测不一定有效果。因此如果我们自己编写代码做这些预处理会非常麻烦，有了建模工具则非常非常简单。

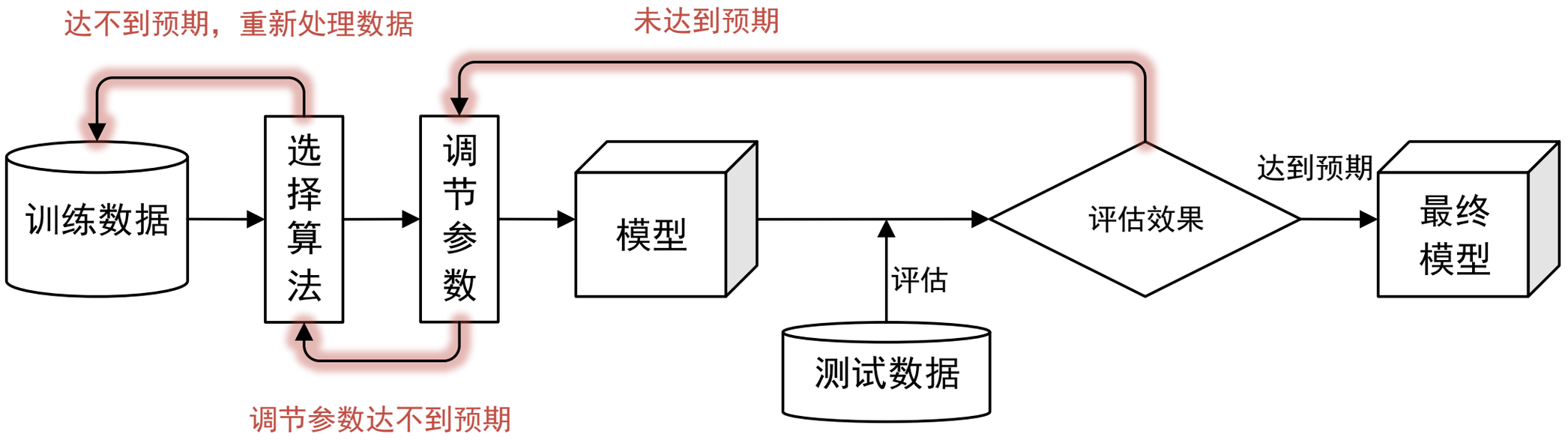
目录 CONTENTS

03

建模与评价

建模与评价

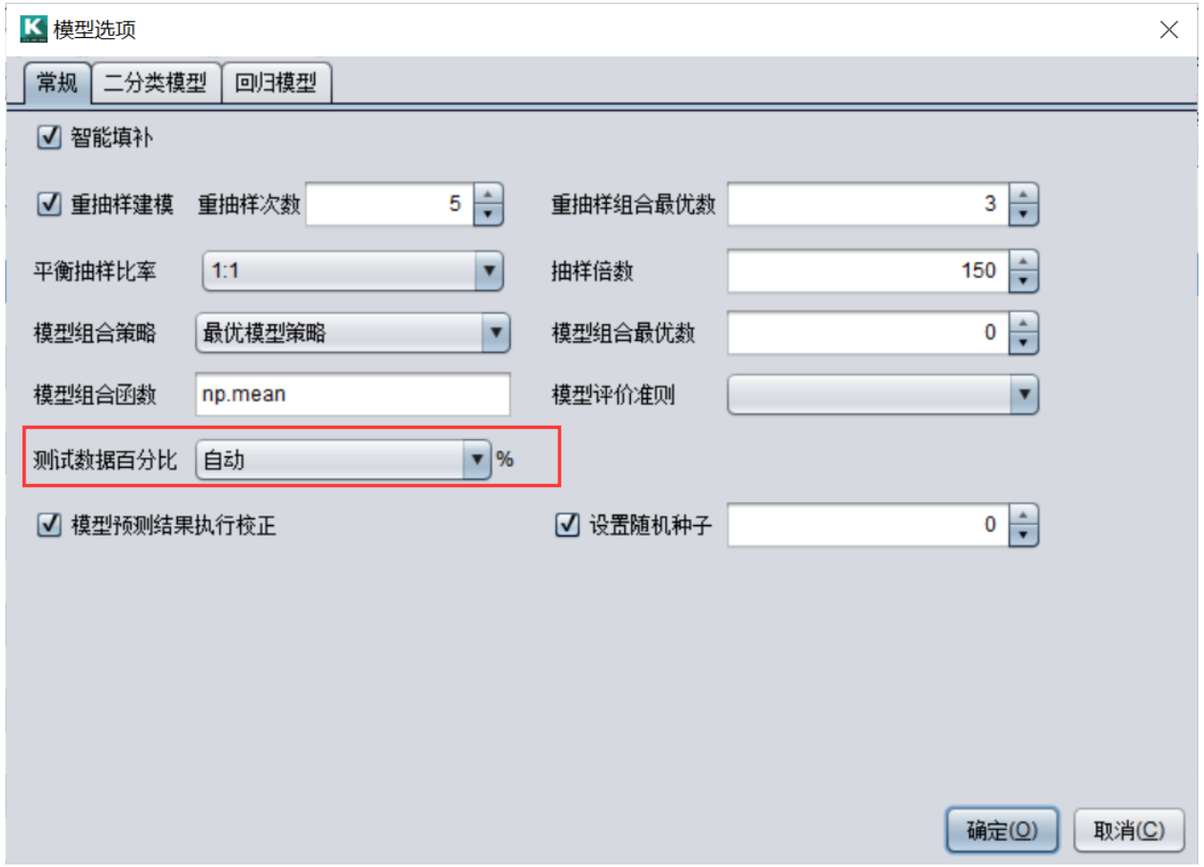
数据预处理后，可以进行建模，建模过程如下图：

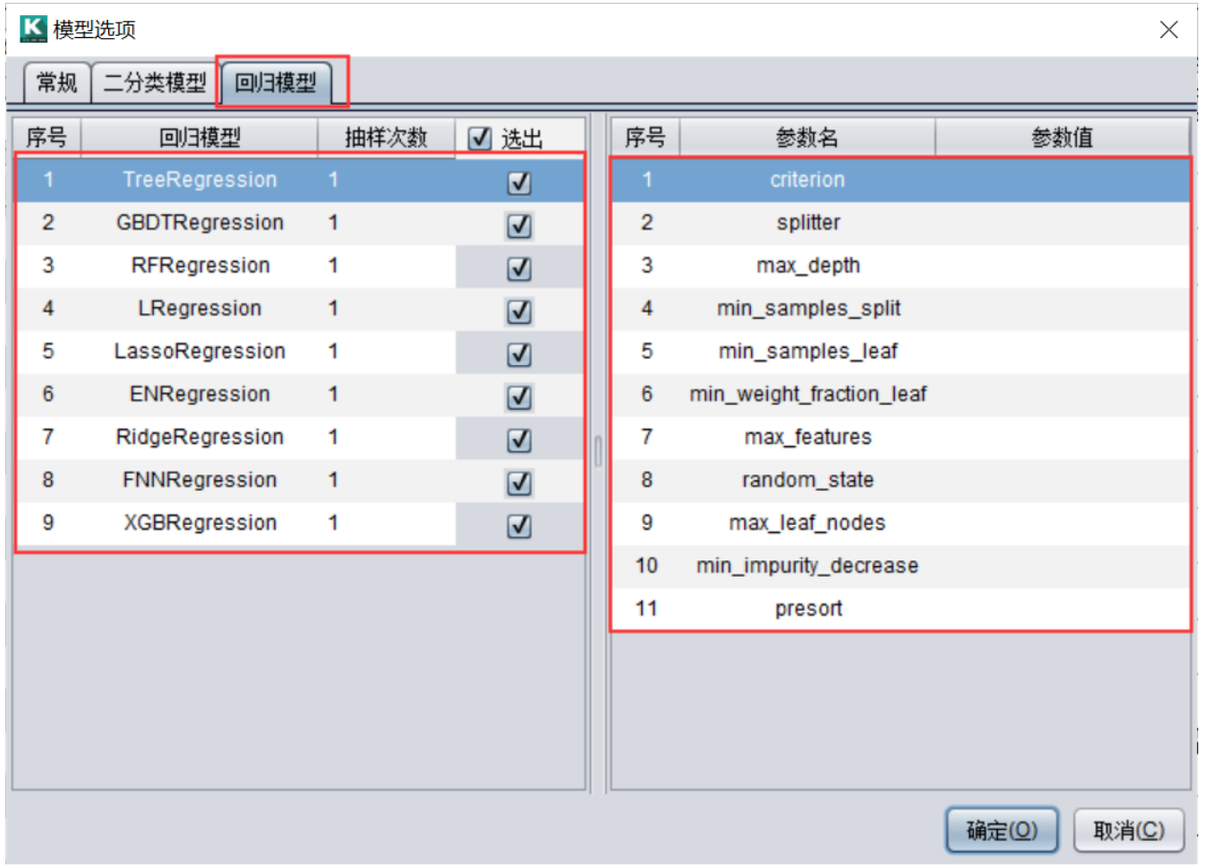


建模的过程是复杂的，交给智能建模工具，它包括了数据探索，数据预处理，建模、评估、模型应用等模块。建模时，它会自动将数据切分成训练集和测试集，并在训练集上建模，测试集上评估。它还包括了参数调节，算法融合等智能化建模手段，用户可以很低成本的建成一个比较理想的模型。

建模与评价

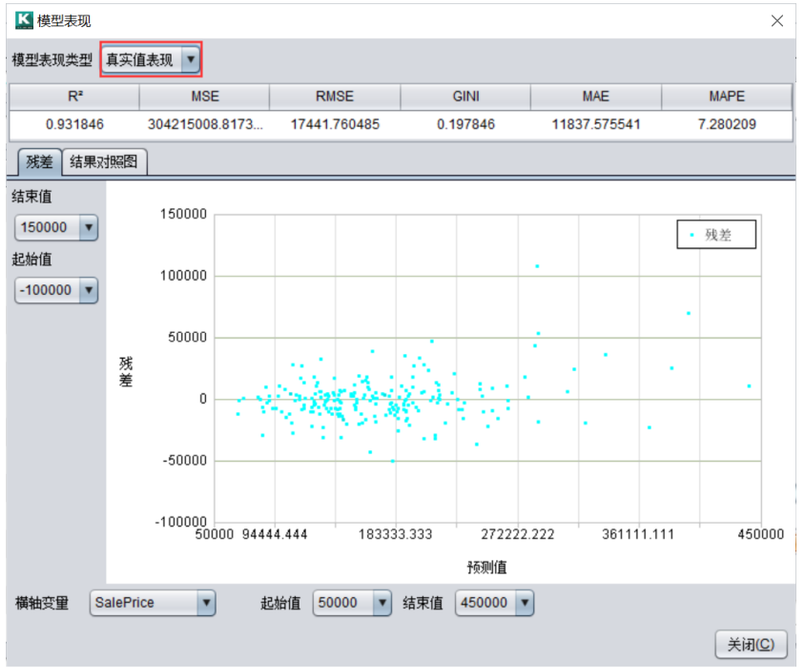
选择模型和调节参数过于复杂，我们只要使用默认的选项就可以了。如下图：



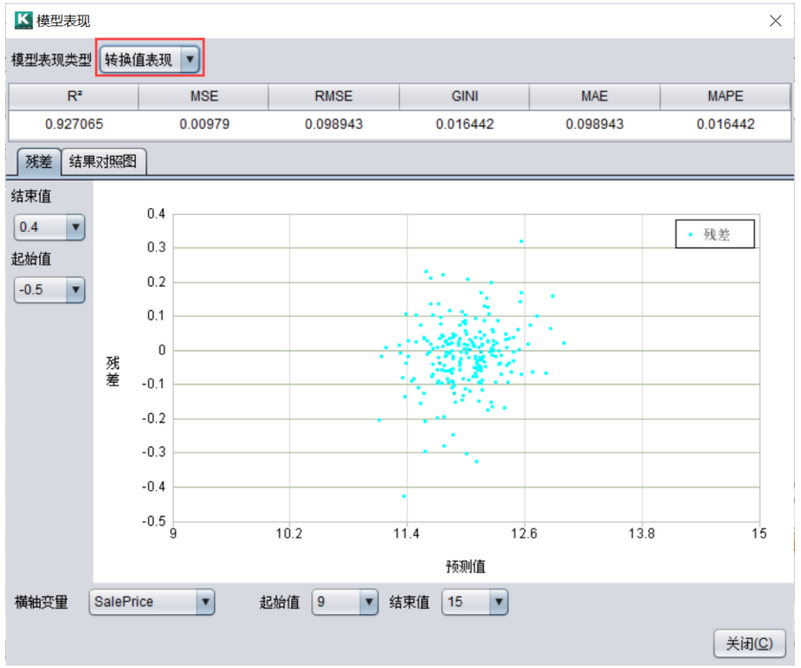


智能建模工具会帮助我们快速的建好模型，采用合适的方法避免过拟合，并在测试集上计算评估指标，方便我们进行评估。

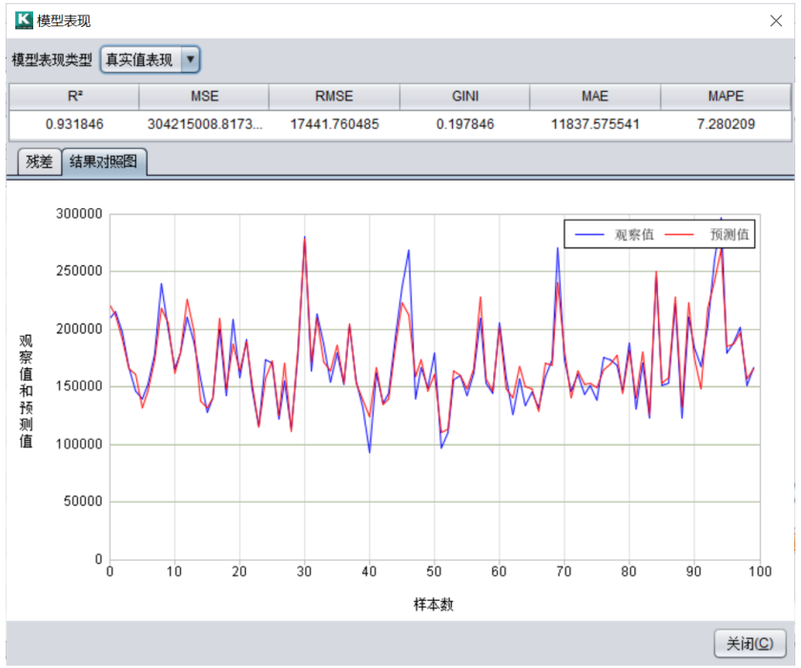
模型表现



真实值表现



转换值表现



结果对照图

从评价指标来看，模型的表现很不错，真实值转换值的R^2都大于0.9，表示模型解释了90%以上的不确定性，真实值的RMSE=17441.76，表示预测的房价与真实房价平均相差17441.76，再看转换的RMSE=0.098943，表示转换后的房价与模型预测的结果相差不到1%，再看残差图，真实值的残差图基本是随机的分布在0周围，特别挑剔的话微微有一点房价越高偏差越大的趋势，但看转换后的残差图则基本是随机的分布在0周围。通过看结果对照图，同样会感受到，预测结果还是比较准确的，基本可以使用。

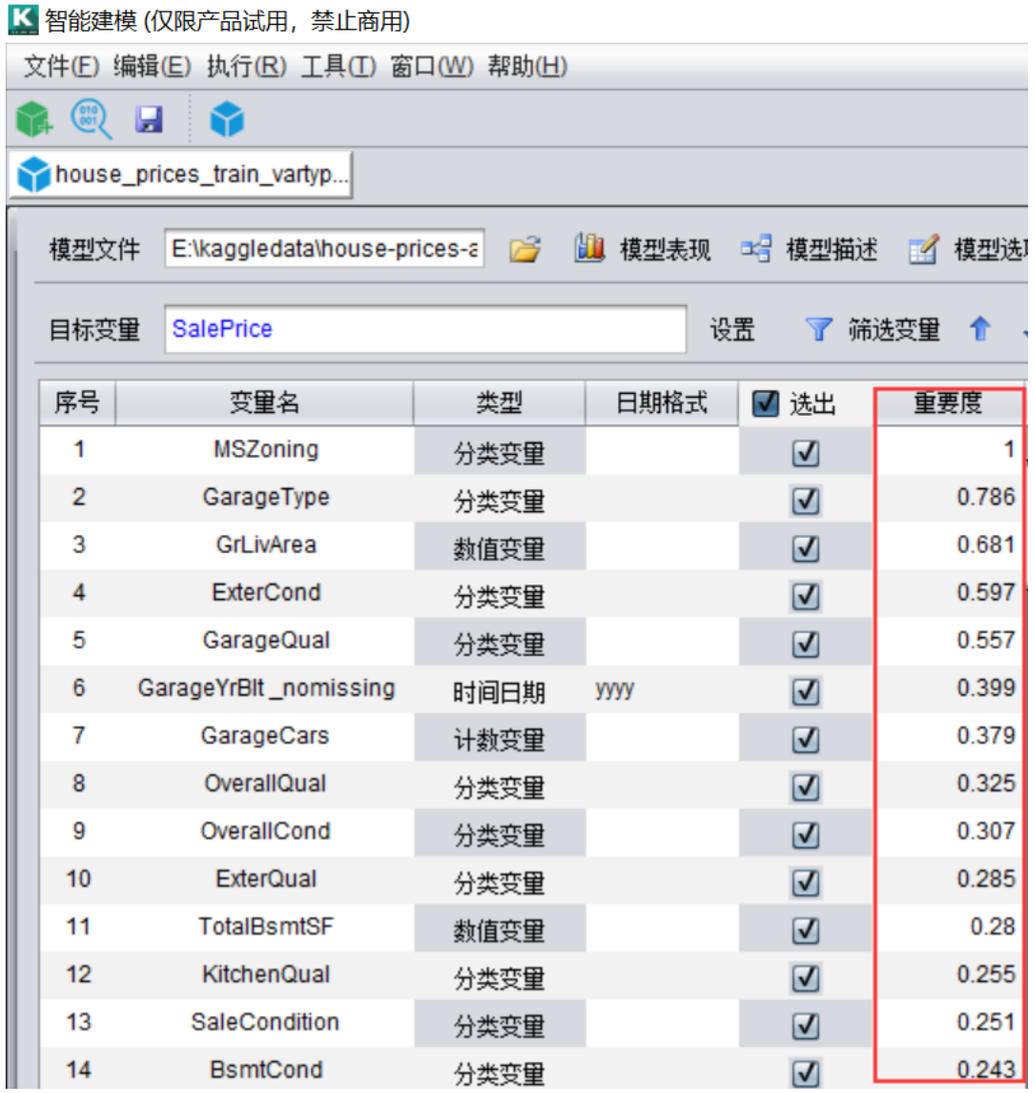
建模使用了哪几个模型：

本次建模选出了Ridge和XGB两个回归模型，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：

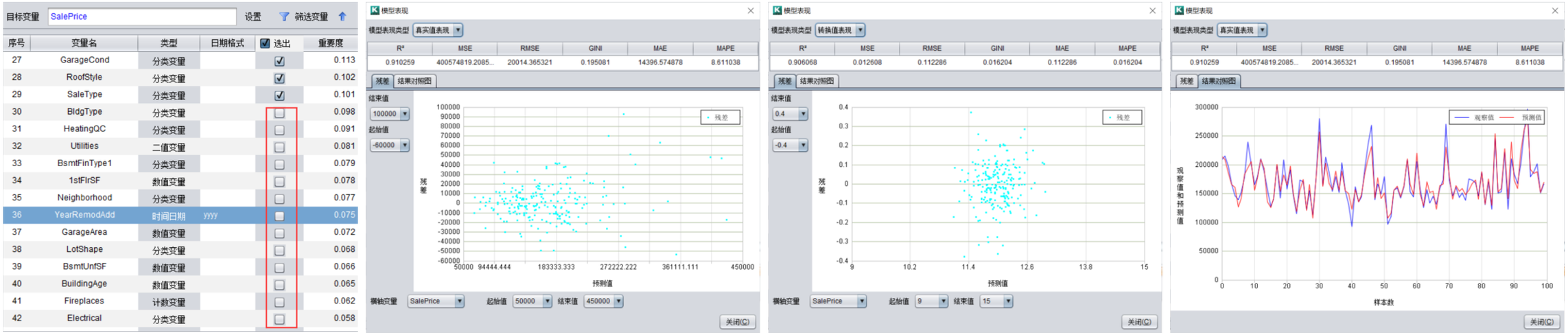
该模型的变量重要度降序排序结果：

MSZoning、GarageType、GirLivArea、ExterCond……其中衍生变量的重要度最高的是GarageYrBit_nomissing第6，其次是LotArea_g第17。BuildingAge、2ndFlrSF_g、GrLivRatio重要度都只要0.065,0.021,0.015，在这我们主要是介绍探索和预处理的方法，在这份数据中这些衍生变量是没有必要的。



序号	变量名	类型	日期格式	选出	重要度
1	MSZoning	分类变量		☑	1
2	GarageType	分类变量		☑	0.786
3	GrLivArea	数值变量		☑	0.681
4	ExterCond	分类变量		☑	0.597
5	GarageQual	分类变量		☑	0.557
6	GarageYrBlt _nomissing	时间日期	yyyy	☑	0.399
7	GarageCars	计数变量		☑	0.379
8	OverallQual	分类变量		☑	0.325
9	OverallCond	分类变量		☑	0.307
10	ExterQual	分类变量		☑	0.285
11	TotalBsmtSF	数值变量		☑	0.28
12	KitchenQual	分类变量		☑	0.255
13	SaleCondition	分类变量		☑	0.251
14	BsmtCond	分类变量		☑	0.243

我们可以使用高重要度衍生变量交互增加新的衍生变量来提高模型表现，感兴趣的同学可以自行尝试，这里我们去除低重要度变量，再建一次模，看看效果如何



这里我们去除了重要度低于0.1的变量，只保留了前29个重要变量，从评价指标到残差图再到结果对照图，模型都没有差太多。因此我们在建模的时候可以选择性的去除一些非重要变量，这样不会影响模型效果，但建模效率会高很多。（此建议不适合竞赛，实际的生产中可以使用。）

目录 CONTENTS

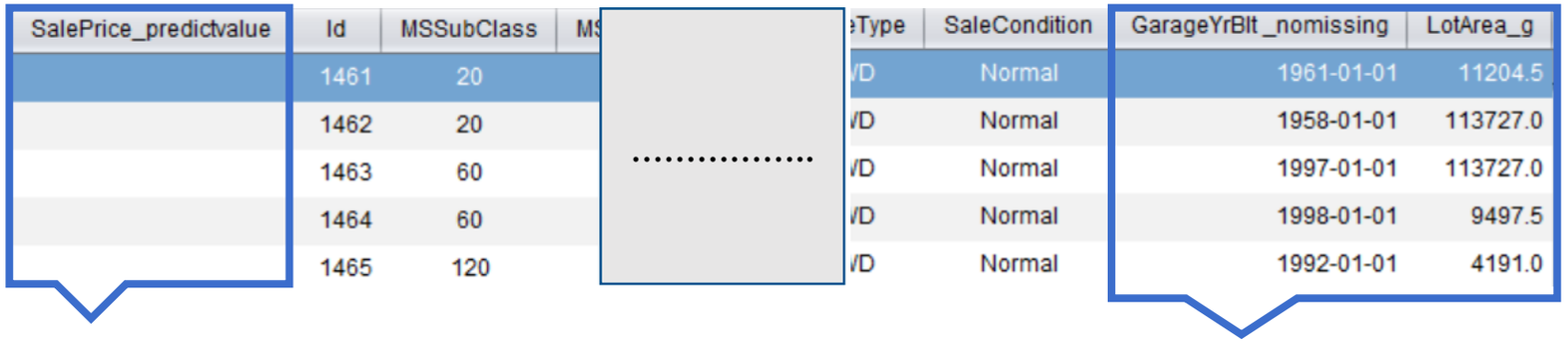
04

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，不需手动处理，如下图：



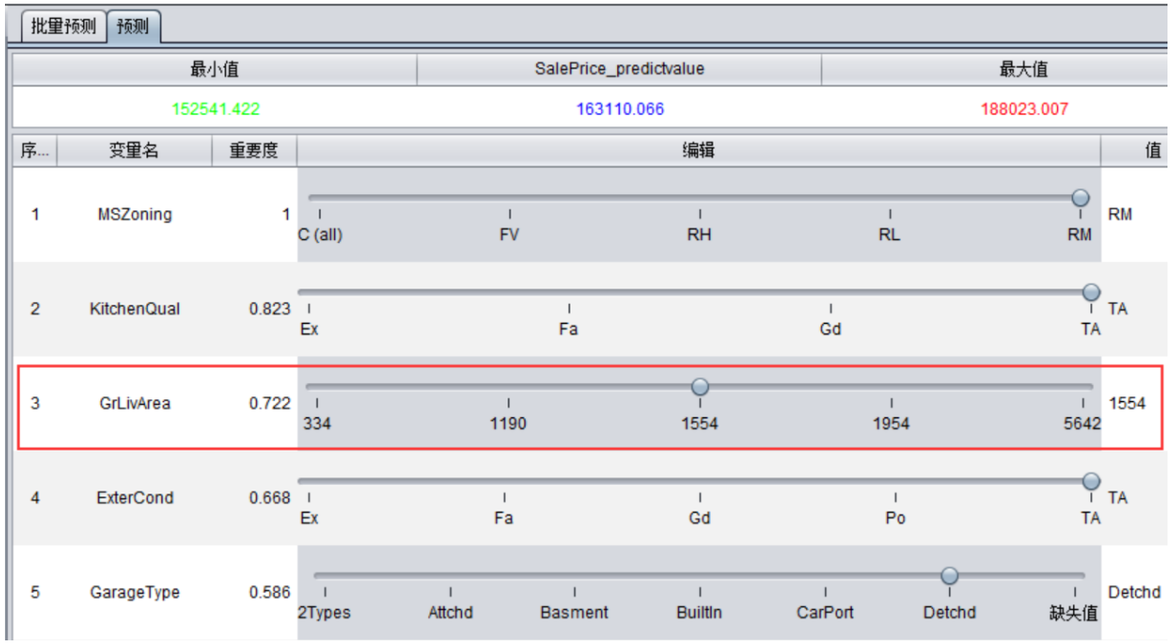
因为我们最后只使用了变量度较高的两个衍生变量，因此待测集只衍生了两个变量。

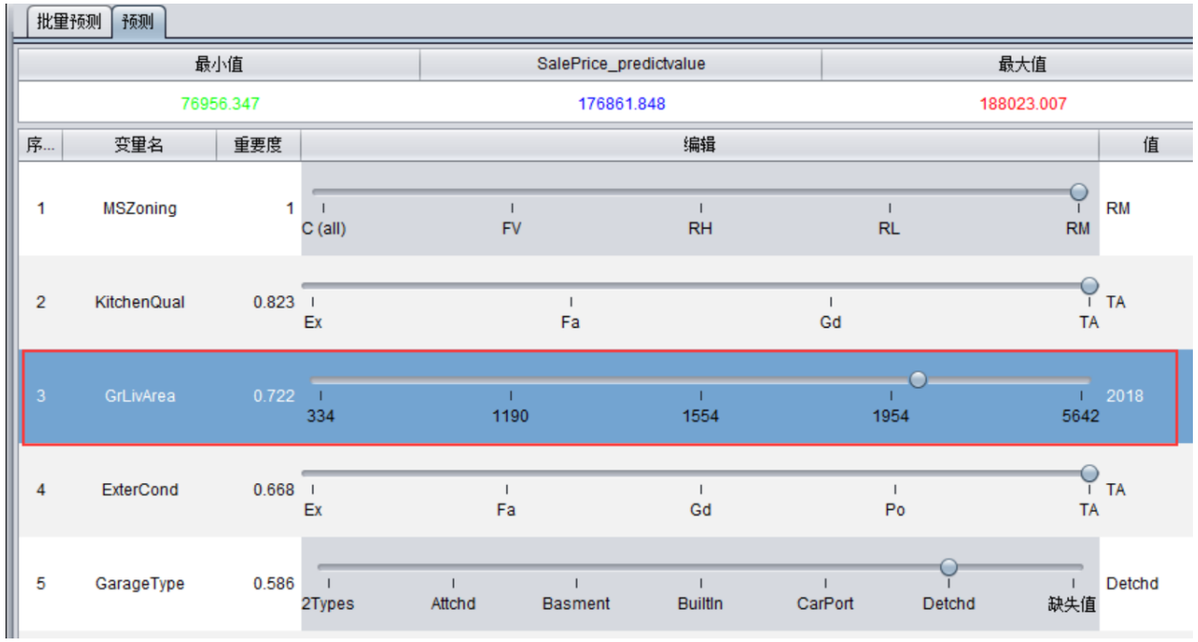
模型预测结果见下图：

SalePrice_predictvalue	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig
118982.109	1461	20	RH	80	11622	Pave		Reg	Lvl	AllPub	Inside
149017.487	1462	20	RL	81	14267	Pave		IR1	Lvl	AllPub	Corner
189376.53	1463	60	RL	74	13830	Pave		IR1	Lvl	AllPub	Inside
183094.941	1464	60	RL	78	9978	Pave		IR1	Lvl	AllPub	Inside
179506.983	1465	120	RL	43	5005	Pave		IR1	HLS	AllPub	Inside

预测结果，房价（SalePrice）的预测值，如第一条预测结果是118982.109，表示该房子的价格大概是118982.109。

我们还可以进行单条预测





如果有人想买房子，其他变量不变的情况下，生活面积GrLivArea为1554时，预测值是163110.066；生活面积GrLivArea为2018时，预测值是176861.848 。据此可以给出客户参考，方便客户决策。

润乾
RAQSOFT

THANKS

创新技术 推动应用进步

www.raqsoft.com.cn